University of Alberta

# Structural Investigations of Ndt80 in complex with DNA: Implications for DNA-protein interactions and transcriptional regulation

by

Jason S. Lamoureux

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Biochemistry

Edmonton, Alberta
Fall 2006

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

*Your file Votre référence*
*ISBN: 978-0-494-23059-6*
*Our file Notre référence*
*ISBN: 978-0-494-23059-6*

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

---

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

**Abstract:**

Ndt80 is the central transcriptional activator of middle genes during sporulation. Upon completion of meiotic recombination, Ndt80 directs the transcription of approximately 150 genes by its association with a specific sequence within the promoter region termed the middle sporulation element (MSE). To gain structural insights into how this protein recognizes its target, we determined the crystal structure of Ndt80 bound to an MSE containing DNA at a resolution of 1.4 Å. This work revealed that Ndt80 is a member of the Ig-fold family of transcription factors, the first example of this family discovered in a non-metazoan organism.

To expand on this work, we did binding studies on Ndt80 and variant MSEs in which single base pairs were mutated to establish the specificity and affinity determinants at each position. These binding studies, in conjunction with the exceptionally detailed structural information available for this complex, allowed us to propose a new mechanism of DNA-protein interactions in which a YpG step is recognized by an arginine side chain through a combination of canonical hydrogen bonds to the major groove of guanine and cation-$\pi$ interactions with the 5' pyrimidine. To determine if this mode of recognition is used by other transcription factors, we searched the database of structures for similar interactions. Remarkably, nearly every major class of transcription factor contains an example of this mode of recognition.

We then undertook a project to solve the crystal structures of Ndt80 bound to the variant MSEs used in the binding studies. We succeeded in

solving 10 of these variants, all with resolutions of 2.0 Å or better. These structures helped to reinforce our model for YpG recognition by arginine residues, but also highlighted the importance of indirect contacts for this type of interaction. Indirect contacts that favor the distorted YpG step, in concert with the arginine interactions, are both critical for the recognition of these steps in Ndt80. Furthermore, these structures provided us with a possible explanation for the preference for a poly-A region rather than alternating A-T sequences. As a result we have structurally supported explanations for specificity at all positions within the MSE. In all, we have solved 12 structures of Ndt80 bound to various DNAs as well as an unbound version making this one of the most comprehensively studied transcription factor systems.

Acknowledgements

I'd like to thank the following people and organizations that have made my time here possible, productive and in many cases enjoyable;

First of all I'd like to thanks my parents, David and Margaret, who always support and encourage me and my siblings, Nadine and Dustin, who are always there for me.

Alberta Heritage Foundation for Medical Research (AHFMR) and the Canadian Institutes of Health Research (CIHR) that funded not only my salary but also the project that I've worked on over the last several years. The Alberta Synchrotron Institute (ASI) which funded several beam-line trips as well as some excellent conferences. In addition, I'd like to thank the administrators that helped me get this funding and keep my program in order, namely Susan Smith and Marion Benedict.

All of the people that contributed to my project; Roger Tsang a summer student over 2 years who helped with EMSA experiments, Jason Maynes whose programming expertise was used to get the PERL script off the ground, Dave Stuart who is, in part, responsible for my project and who contributed genetic data to my first paper with the help of one of his students Cynthia Wu.

Members of the Glover lab that taught me all the practical aspects of lab work namely Ruth Green and Scott Williams and Ross Edwards for his expertise with computers and software.

People around the department that made late nights "at the lab" and time away from the lab more enjoyable; Alex, Scott, Sheetal, Jody, Trevor, Ross, Donx, Megan, Steve, Nina, Kaari, Magnus, Dave H., and all the noobies.

Finally I'd like to thank my supervisor, Mark Glover, whose decision to take on a frantic undergrad looking for a project supervisor resulted in one of the longest undergrad research projects ever. His guidance and support were invaluable to these early stages of my career. Any success I have enjoyed in science can be traced back to him - any errors, however, are my own.

TABLE OF CONTENTS

**Chapter 1: Introduction**

**NDT80 and meiosis**.....2

Purpose of meiotic cells.....2

Role of Ndt80 in sporulation.....4

Middle sporulation element.....8

Post transcriptional control of Ndt80.....10

Repression at MSE sites.....10

**Protein-DNA interactions - a general view**.....12

Modes of protein-DNA recognition.....12

Example of the complexity of DNA-protein interactions.....15

Ndt80 as a model system.....20

**Research Overview**.....23

**References**.....25

**Chapter 2: The crystal structures of the core domain of Ndt80 and the DNA - binding domain of Ndt80 bound to DNA**

**Summary**.....30

**Introduction**.....31

**Experimental procedures**.....32

Cloning and vector construction.....32

Protein expression and purification.....32

Electrophoretic mobility shift assays (EMSA).....35

Proteolytic mapping.....36

DNA purification.....36

Crystallization and data collection.....40

Structure determination and refinement.....44

**Results and discussion**.....47

Elucidation of the Ndt80 DNA-binding domain.....47

Binding affinity of Ndt80 for mutant MSEs.....47

Overall structure....51
Comparison with other Ig-fold transcription factors....56
Ndt80-MSE specificity: 5'-YpG-3' recognition....58
Ndt80-MSE specificity: minor groove recognition....62
Ndt80-MSE specificity: backbone contacts....65
Ndt80 flexibility and DNA binding....67
Interactions between Ndt80 and other transcription factors....67
**References**....70

**Chapter 3: A new mode of 5'-YpG-3' step recognition by amino acids: Coupled hydrogen bond and stacking interactions**

**Summary**....77
**Introduction**....78
**Experimental procedures**....81
Parsing the database....81
Assessing unstacking....81
Aligning to a reference DNA....83
DNA helical parameter calculations....84
**Results and discussion**....85
Database search for interactions between arginine and 5'-YpG-3'....85
Conformational analysis of unstacked 5'-YpG-3' steps....86
Arginine - 5'-YpG-3' interactions and sequence-specific recognition....91
Evidence for arginine-induced distortion of YpG steps....96
5'-TpG-3' vs. 5'-CpG-3' recognition....97
Histidine - 5'-YpG-3' recognition....99
Generality of protein induced unstacking of YpG steps....100
**References**....102

**Chapter 4: Principles of protein-DNA recognition revealed in the structural analysis of Ndt80-MSE DNA complexes**


**Summary**.....111

**Introduction**.....112

**Experimental procedures**.....114

Crystallization and Data Collection.....114

Structure Determination, Refinement and Analysis.....116

**Results**.....119

Variants.....123

vG1A/A9T.....123

vG1C.....126

vA4G.....126

Mutants.....128

mA4T and mA6T.....128

mC5T.....130

mA7T.....133

mA8T.....135

mA7G.....137

mA9C.....139

**Discussion**.....142

GC-rich region recognition.....143

Poly-A tract recognition.....145

Implications for modeling meiotic transcriptional activation in S. cerevisiae.....148

**References**.....151

**Chapter 5: General Discussions and Conclusions**

**Ig-fold transcription factors**.....156

Structural similarities.....156

The possibility of evolutionary links.....158

**Transcriptional regulation of meiosis**.....159

**DNA-protein recognition principles**.....162

On "recognition codes".....163

Technical Issues of DNA modelling.....164

**Conclusions**.....165

**References**.....166

**Appendix A: List of PDB used in Chapter 4 as representative of the DNA-protein complex database**.....169

**Appendix B: Perl Script used to parse the database (Appendix A)**.....173

**List of Tables**

**Table 2.1** Summary of X-ray experiments of Ndt80(59-340) and Ndt80(1-340)-MSE complex.....43

**Table 3.1** Summary of database search and DNA parameters for YpG steps contacted by arginines.....89

**Table 3.2** Summary of database search and DNA parameters for YpG steps contacted by histidines.....90

**Table 4.1** Comparison of previously reported binding and activation data and the RMSD of the variant structures in comparison to wild type.....115

**Table 4.2** Summary of X-ray experiments of variant and mutant Ndt80-MSE complexes.....118

**List of Figures**


**Figure 1.1** The central role of Ndt80 as a transcriptional regulator of meiosis....7

**Figure 1.2** Examples of DNA-protein interactions....14

**Figure 1.3** Characteristic interactions of a Zif268-like zinc finger and its DNA subsite....16

**Figure 1.4** The architecture of the 3 zinc fingers in the Zif268 transcription factor....17

**Figure 1.5** Similar DNA substrates of other Ig fold transcription factors....22

**Figure 2.1** Expression, purification and crystallization of Ndt80(59-340) and Ndt80(1-340)....34

**Figure 2.2** Limited proteolysis of Ndt80 shows a region on the N-terminus that is stabilized upon the addition of MSE containing DNA..38

**Figure 2.3** DNA purification scheme....39

**Figure 2.4** Crystals and diffraction from Ndt80(59-340)....41

**Figure 2.5** Figure 2.5 Ndt80 (1-340) MSE complex....42

**Figure 2.6** Sample electron density....46

**Figure 2.7** Defining the minimal DNA binding domain of Ndt80....48

**Figure 2.8** Mutation of the MSE reduces Ndt80-binding affinity....50

**Figure 2.9** Model of Ndt80(1-340) bound to DNA....52

**Figure 2.10** Topology diagram of Ndt80 DNA binding domain....53

**Figure 2.11** Overview of Ndt80(1-340)-MSE complex....55

**Figure 2.12** Ndt80-DNA interface in the major groove....59

**Figure 2.13** 5'-YpG-3' Recognition by Ndt80....60

**Figure 2.14** Ndt80-DNA interface in the minor groove....63

**Figure 2.15** Schematic of Ndt80-DNA interactions....66

**Figure 3.1** Alignment of NDT80-DNA complex with reference DNA structure....80

**Figure 3.2** Criteria for identifying unstacked 5'-YpG-3' steps....82

**Figure 3.3** Summary of base displacements....87

**Figure 3.4** Evidence for arginine-induced unstacking 5'-YpG-3' steps....97
**Figure 4.1** Overall diagram of the Ndt80-DNA complex.....120
**Figure 4.2** Major groove view of the wild type structure.....121
**Figure 4.3** Minor groove view of the wild type structure.....122
**Figure 4.4** Structural rearrangements seen in the vG1A/A9T complex.124
**Figure 4.5** Structural rearrangements seen in the vA4G complex.....127
**Figure 4.6.** Mutation of the pyrimidines in the 5'-YpG-3' steps.....129
**Figure 4.7.** Stereo view of the changes in the mC5T structure.....132
**Figure 4.8** Structural rearrangements of the mA7T mutant.....134
**Figure 4.9** Structural rearrangements of the mA8T mutant.....136
**Figure 4.10** Structural consequences of the mA7G mutation.....138
**Figure 4.11** Structural consequences of the mA9C mutation.....141
**Figure 5.1** Comparison of Ig-fold transcription factors.....157
**Figure 5.2** Transcriptional activation versus active Sum1 concentrations in a model system.....161

**List of Abbreviations**

| | |
|---|---|
| Å | Angstroms ($10^{-10}$ meters) |
| aa | amino acid |
| APS | Advanced Photon Source |
| βME | β-mercaptoethanol |
| BI DNA | sub-type of B form DNA characterized by $\epsilon$ and $\zeta$ torsion angles in the trans and gauche minus range respectively, ($\epsilon$ - $\zeta$ ≈-90°) |
| BII DNA | sub-type of B form DNA characterized by $\epsilon$ and $\zeta$ torsion angles in the gauche minus and trans range respectively, ($\epsilon$ - $\zeta$ ≈+90°) |
| BTP | bis-tris-propane |
| DNA | deoxyribonucleic acid |
| DTT | dithiothreitol |
| EDTA | (ethylenedinitrilo)-tetraacetic acid |
| EMSA | electro-mobility shift assay |
| \|F\| | diffracted X-ray structure factor amplitude |
| GST | glutathione-S-transferase |
| kDa | kiloDalton |
| Ig | immonuglobulin |
| IPTG | isopropyl β–D-thiogalactopyranoside |
| λ | wavelength |
| LB | Luria Bertani |
| MAD | multiwavelength anomolous dispersion |
| MALDI-TOF | matrix assisted laser desorption ionization - time of flight |
| MES | (2-[N-morpholino]ethanesulfonic acid) |
| NMR | nuclear magnetic resonance |
| PAGE | polyacrylamide gel electrophoresis |
| PBS | phosphate buffered saline |
| PCR | polymerase chain reaction |
| PDB | protein data bank |
| PEG | polyethylene glycol |

| | |
|---|---|
| PMSF | phenylmethyl sulfonyl fluoride |
| R | purine (i.e. either adenine or guanine) |
| RMSD | root mean square deviation |
| SDS | sodium dodecylsulfate |
| Tris | tris(hydroxymethyl)aminomethane |
| Y | pyrimidine (i.e. either thymine or cytosine) |

# Chapter 1:
# Introduction

## NDT80 and meiosis

### Purpose of meiotic cells

Meiosis is a unique process in biology, not only as a prerequisite for sexual reproduction, but also it introduces variability to the organism. While many other cell processes are designed to maintain an equilibrium, meiosis shuffles the genome through recombination of the homologous chromosomes. Meiosis can break up favorable gene combinations and it requires large expenditures of time and energy (Bell, 1982); but what benefit does it serve the organism? This leads to the question asked by many evolutionary biologists: why did meiosis and sexual reproduction evolve and how has it persisted? With little supporting empirical evidence, the prevailing theory is that the process of evolution is dependent on and accelerated by meiosis to create new arrangements of genes (or at least alleles of these genes). Recently a paper has verified the validity of this theory by demonstrating that yeast that reproduce sexually adapt to a harsh new environment more rapidly than their asexually reproducing counterparts (Goddard et al., 2005).

Essentially meiosis is akin to a poker game with chromosomes. Most of the time this shuffling results in minor changes that accounts for the variation we see in our own individuality, but occasionally a royal flush is dealt and the organism improves upon or gains an ability not seen in its peers. Conversely and more likely the opposite can occur and result in detrimental changes. So why did nature develop such a risky process if the most likely outcome is

destructive to the organisms survival. The reason appears to for the collective good. Although individuals may have detrimental changes, the rare positive change ensures that the species can remain competitive within their ever changing environment. In addition, there is typically competition for resources amongst species and if one gains an ability that allows it to out compete the other species then the others must adapt or perish. Despite its necessity and potential gain, the actual process of meiosis is fraught with danger.

Meiotic cells, regardless of the organism, face two major tasks. The first is to ensure proper division of the genetic material among the haploid progeny. Meiosis consists of extensive recombination of the homologous chromosomes that must be properly resolved before the cells commence the subsequent two rounds of meiotic division. Failure to do so typically results in unviable progeny. For example, in humans, failure of chromosome 21 to segregate leads to Down's syndrome (Lejeune et al., 1959), which is a relatively mild trisomy syndrome in that it does not result in death at an early stage of the embryo. However, this "mild" trisomy results in varying degrees of mental retardation, shortened life expectancy and health issues; all extreme issues as far as an individual is concerned (Cavalli et al., 2003).

The second task of meiotic cells is to provide a good home for the haploid genome created by the meiotic divisions. The morphogenesis of these cells is organism specific and designed for the sexual reproductive methods of that organism. In higher organisms there is a distinction between the male and female of the species, such as sperm/oocyte in humans. But in

*Saccharomyces cerevisiae* the meiotic nuclei's home is an ascospore in both mating types. These two processes of meiosis and gametogenesis, although distinct, must clearly be coupled both spatially and temporally to result in viable gametes.

**Role of Ndt80 in sporulation**

In *Saccharomyces cerevisiae*, the process that couples the replication and segregation of chromosomes with morphological changes that accompany the development of the gamete is termed sporulation (Roeder, 1997). This process was proposed to be regulated by a transcription cascade, temporally regulated by at least four distinct waves of gene expression: early, middle, mid-late, and late (Mitchell, 1994). At this time the early gene promoters were the most studied and it was noted that they contained similar regulatory elements. It was proposed that other temporal class genes may contain distinct regulatory elements that could initiate each subsequent phase in sporulation when given the correct queues.

Sporulation can be initiated in diploid yeast cells by introducing starvation conditions (nitrogen starvation in the presence of a non-fermentable carbon source). Shortly after these conditions are met, typically by a change in growth media, the early wave of sporulation specific gene transcripts are seen (reviewed by (Kupiec et al., 1997)). Many early genes, such as *HOP1* and *DMC1*, encode for proteins with roles in chromosomal synapsis and recombination (Kupiec et al., 1997). In 1988 an Inducer of

MEiosis (IME1) was discovered (Kassir et al., 1988). Subsequent research revealed that IME1 is a transcriptional activator that with DNA binding protein UME6 is required for the transcription of early genes (Kassir et al., 2003). The discovery of this key early gene activator spurred on the search for a key middle gene activator in this cascade.

Throughout the early phase of meiosis the paired homologous chromosomes are monitored for unresolved recombination intermediates. Until these intermediates are resolved the cells will halt at the pachytene stage of meiosis I. The successful resolution of the recombination intermediates is so important that the pachytene stage of meiosis accounts for over half the time required for both meiosis I and II in mammals (7 out of 12 days for male mice), although the length of this stage varies between the sexes and among species (Alberts et al., 1994). Once recombination is complete and the chromosomes are ready for segregation, there is a transition from prophase to metaphase I. In both yeast and metazoans this transition coincides with the beginning of middle gene expression and requires a 'maturation promoting factor' (MPF). In yeast the MPF is composed of CDC28, the catalytic subunit, and one of several B-type cyclins, which act as a regulatory subunit. An early study, using a temperature sensitive *cdc28* allele, demonstrated that this mutant can enter meiosis but arrests at pachytene, shortly before the prophase to metaphase I transition (Shuster and Byers, 1989). These arrested cells have duplicated spindle pole bodies and fully synapsed chromosomes. It was the similarity of phenotypes between

this *cdc28* mutant and another mutation, *ndt80-1*, that a candidate for control of middle gene expression was discovered.

In 1995, Xu et al. used a dityrosine fluorescence assay to screen for mutants defective in spore formation found a mutation *ndt80-1* (for non dityrosine) (Xu et al., 1995). This particular mutation, like *cdc28*, resulted in the arrest of development in the pachytene stage of meiosis. Furthermore, the behaviour of *ndt80/spo11* or *ndt80/rad50* double mutants suggested that *NDT80* is not required for recombination since neither the *spo11* nor the *rad50* mutations, which cause a bypass of meiotic recombination, alleviated *ndt80*-dependent arrest of meiosis. In addition, a *dmc1* mutation, which blocks completion of meiotic recombination and thereby leads to activation of the recombination checkpoint and pachytene arrest (Lydall et al., 1996), blocks *NDT80*-activated gene expression. At this point, the precise role of Ndt80 was unknown. It was more than two years after this work that Chu et al. demonstrated that Ndt80 is a transcription factor that directly controls the expression of many middle sporulation genes via a recognition element in the promoter region of these genes (Chu and Herskowitz, 1998).

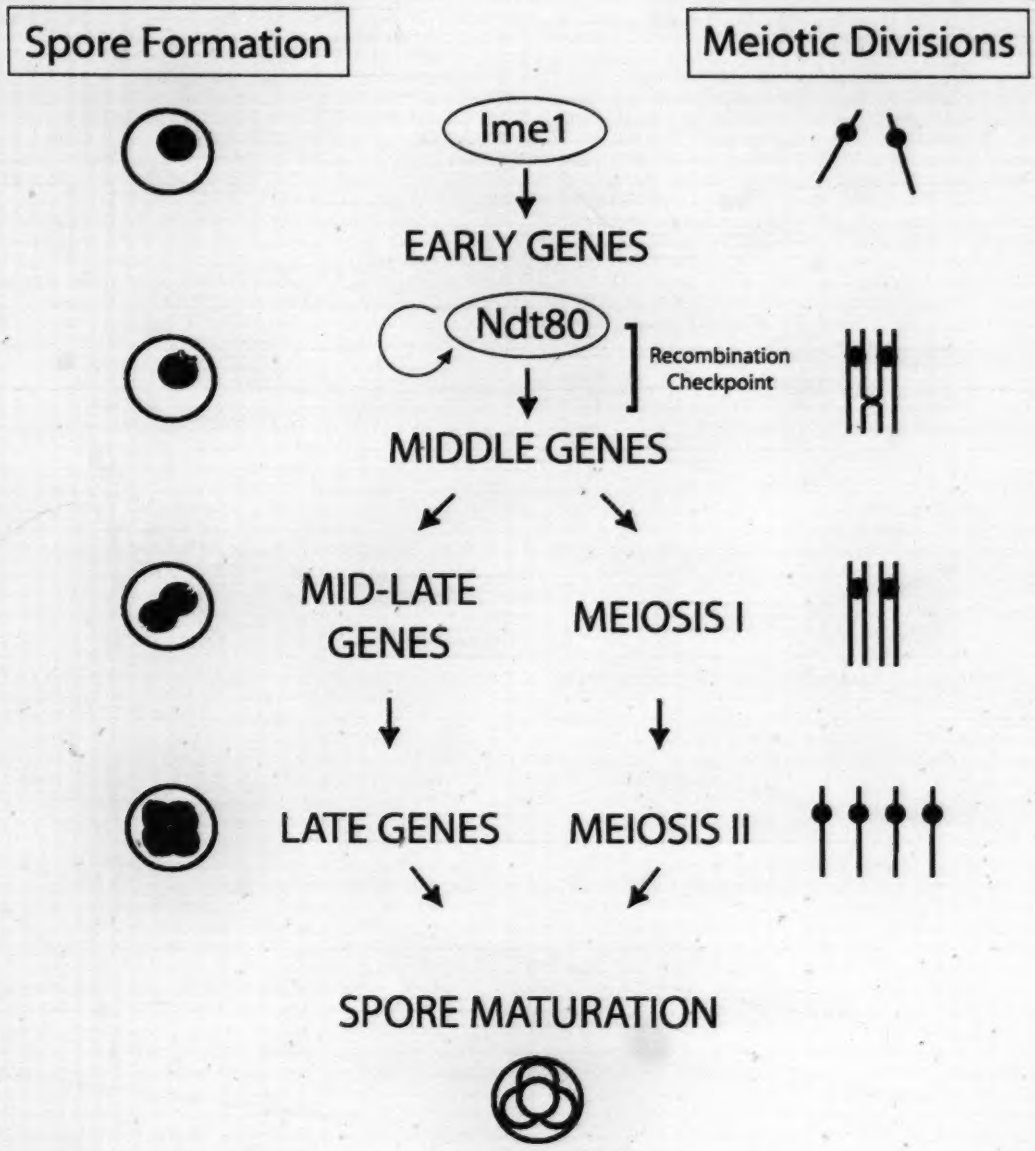


**Figure 1.1 The central role of Ndt80 as a transcriptional regulator of meiosis.** The diagram includes landmark events in sporulation. On the left are diagrams illustrating the morphological changes involved in spore formation and on the right are the coordinated events of meiotic divisions. Once the recombination checkpoint is passed, the cells are committed to meiosis, prior to that they can resume vegetative growth if conditions allow. Adapted from (Chu et al., 1998; Chu and Herskowitz, 1998).

**Middle sporulation element (MSE)**

Remarkably, the DNA sequence that is recognized by Ndt80 was discovered independently of Ndt80 research as a new sporulation control element (Hepworth et al., 1995; Ozsarac et al., 1997). Hepworth et al. described a 15 base pair sequence in the 5' upstream region of the middle gene SPS4 that conferred sporulation specific expression of a reporter gene. They further noted that this sequence was similar to one that is critical for the sporulation specific expression of SPR2 and proposed that it may by a general regulatory element of middle genes. This work was verified by Ozsarac et al., who identified a related sequence in the promoter region of SPR3 and formulated a consensus based on the sequences of other middle genes, resulting in a 9 base pair consensus termed the middle sporulation element (MSE). Electrophoretic mobility shift assays (EMSA) using ectopically expressed Ndt80 from yeast extracts, or recombinant Ndt80 purified from *E. coli*, have demonstrated that Ndt80 specifically binds this regulatory sequence (Chu and Herskowitz, 1998).

Compelling evidence for the central role of Ndt80 in middle sporulation was provided by DNA microarray experiments of the transcriptional program during sporulation (Chu et al., 1998). In addition, there was a special focus on NDT80 including ectopic expression of this gene in vegetative cells and elimination of NDT80 during sporulation. These experiments showed over 150 genes that were strongly induced during the middle phase of sporulation with 70% of these having a possible MSE immediately upstream. When

Ndt80 is ectopically expressed, over 200 genes are induced at least 3 fold. Of these induced genes, 42% were expressed in the middle phase of sporulation with fewer than 20% exhibiting the metabolic, early, early-mid, mid-late or late expression profiles. Also 62% of the genes ectopically induced by Ndt80 contain at least one MSE in their promoter regions. Thus there is a strong, although not absolute, correlation between the presence of an MSE, a middle sporulation expression profile, and the ability to be induced by Ndt80.

While early genes appear to be primarily involved in the early meiotic portion of sporulation, middle genes begin to demonstrate the coordinated control of both meiosis and gametogenesis. Many of the genes activated by Ndt80 are required for the disassembly of the synaptonemal complex, such as UBC9, which allows the physical separation of chromosomes to occur. Also, factors required for further progression through the developmental process, such as several of the B-type cyclins (5 of the 6 CLBs) show an increase in expression. Finally, genes involved in spore morphogenesis, such as SPO20 required for prospore membrane formation, begin to be expressed. The integration of these two distinct processes highlights the importance of this cascade. Both the phenotype of the *ndt80* mutant as well as the critical nature of integration suggest activation of Ndt80 may be under checkpoint control. Furthermore, in *dmc1* mutants, which do not proceed through the meiotic recombination checkpoint, middle gene expression is completely blocked, in spite of the fact that Ndt80 is still transcribed (Chu and Herskowitz, 1998; Hepworth et al., 1998). If however RAD17, a checkpoint gene, is

mutated as well, middle gene expression is restored to the *dmc1/rad17* cells, suggesting that the meiotic recombination checkpoint regulates activation of middle sporulation gene expression.

**Post transcriptional control of Ndt80**

It has been suggested that the checkpoint dependent regulation of *NDT80* is post-transcriptional (Chu and Herskowitz, 1998; Tung et al., 2000). Support for this idea comes from the finding that Ndt80 is extensively phosphorylated during meiosis and that this modification is required for the activation of middle sporulation genes. Inhibition of this phosphorylation is proposed to be one mechanism to stimulate checkpoint-induced arrest at pachytene. Although the mechanism by which this checkpoint activation is propagated is unknown, Ndt80 appears to exist in various phosphorylated states. In addition, Ndt80 binds with high affinity to the MSE in a non-phosphorylated state *in vitro*, suggesting that phosphorylation may be involved in either inhibiting activation or sequestering Ndt80 away from its MSE targets.

**Repression at MSE sites**

Some MSE sequences have been shown not only to activate middle genes, but also to repress the transcription of these same genes during vegetative growth (Pierce et al., 2003; Pierce et al., 1998). MSE-mediated repression requires Sum1, a DNA binding protein that binds a subset of the

MSEs and recruits the Hst1 histone deacetylase to these genes, presumably to promote an inactive chromatin structure at these genes during vegetative growth and early sporulation (Xie et al., 1999). These findings lead to the suggestion that the relative affinities of an MSE for either Sum1 or Ndt80 could govern the precise timing and level of activation of a gene under its control (Xie et al., 1999; Lindgren et al., 2000). Consistent with this model, the levels of Sum1 protein drop during the mid-phase of sporulation, at a time when Ndt80 levels rise and the middle sporulation genes are activated (Chu et al., 1998; Lindgren et al., 2000). Additional work demonstrated that Ndt80 and Sum1 have overlapping but distinct binding requirements and extended MSE consensus sequences have been proposed for both proteins (Pierce et al., 2003). It has also been shown that the binding of either Ndt80 or Sum1 to an MSE is mutually exclusive *in vitro*. Research on Ndt80 and Sum1 was also done *in silico* using a computer model to probe the transcriptional network of these proteins (Wang et al., 2005). They proposed that competition for binding sites between Ndt80 and Sum1, in addition to positive autofeedback of Ndt80, can account for the sharp and precise temporal control of middle genes. They further proposed that these two features of activator/repressor competition and positive autofeedback may be general features of systems where such sharp temporal boundaries of gene expression are required.

## Protein-DNA interactions - a general view

### Modes of protein-DNA recognition

Since the structure of DNA was proposed by Watson and Crick (Watson and Crick, 1953), scientists have been trying to understand how DNA binding proteins recognize their DNA substrates. Early theoretical work by Seeman *et al.* suggested that sequence specific interactions could be obtained through interactions of the protein side-chains with the face of the DNA base-pairs exposed in the major groove; so called "direct readout" (Figure 1.2A) (Seeman et al., 1976). For example, the guanidinium group of an arginine can hydrogen bond to the O6 and N7 atoms of a guanine base. Similarily, asparagine and glutamine side chains can recognize the N6 and N7 atoms of an adenosine. The first high resolution structures of protein-DNA complexes in the late 1980s, such as Trp repressor and phage 434 repressor, validated this theory (Aggarwal et al., 1988; Otwinowski et al., 1988). However, these early structures also suggested that another mode of recognition was being employed. It was proposed that sequence specific variations in the structure of the DNA double helix could be recognized by the protein through a type of interaction termed "indirect readout" (Figure 1.2B) (Aggarwal et al., 1988; Anderson et al., 1987; Otwinowski et al., 1988), one of the best such examples is the TATA-box binding protein (TBP). This protein utilizes the high degree of flexibility of alternating TA sequences to severely bend the DNA toward the major groove. This creates a wide and shallow minor groove that is recognized by the central anti-parallel β sheet that forms

a primarily hydrophobic interface with the DNA. Indirect readout involves the overall recognition of the deformed DNA including the phosphate backbone contacts. Direct readout is also involved in the hydrophobic interface, but its role seems less essential than the indirect recognition in this case. Although this is an extreme example, indirect readout also covers the more subtle and poorly defined examples where a protein's apparent nucleotide sequence preference is not ascribed to a direct contact. That is, any recognition that is not mediated through the distinct chemical composition of the different bases is typically termed indirect readout.

Since these two modes of recognition were described, hundreds of protein-DNA complexes have been solved through NMR and X-ray crystallographic methods (http://ndbserver.rutgers.edu listed 1009 structures containing both DNA and protein). Yet, it has still been difficult to determine fundamental principles that can be used to predict DNA binding preferences of proteins *a priori*.

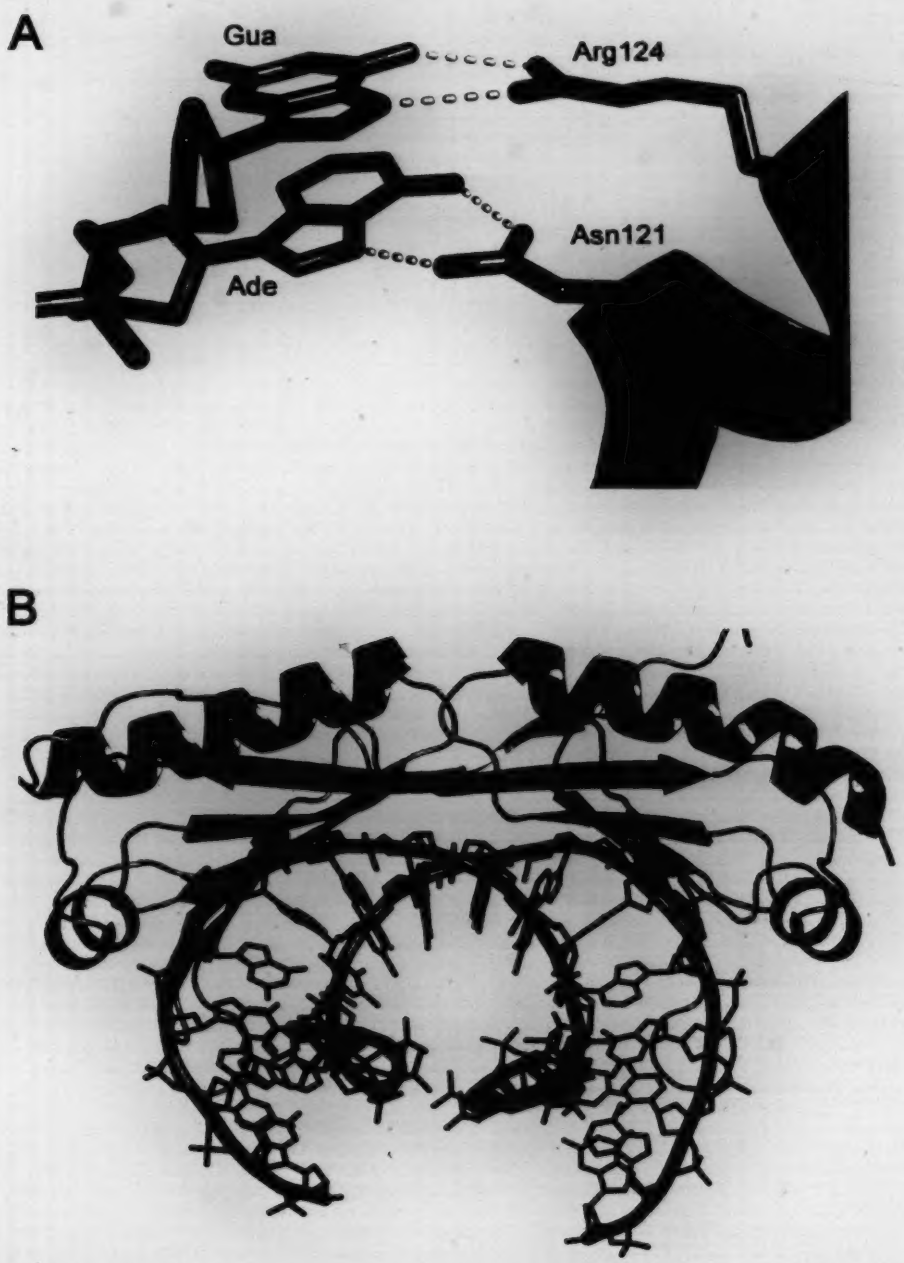


**Figure 1.2 Examples of DNA-protein interactions.** **(A)** Direct readout of bases in the major groove by hydrogen bonds to amino acid side chains. Shown is guanine and adenine recognition by arginine and asparagine side chains respectively in a Zif268 variant complex (Elrod-Erickson et al., 1998). (B) Indirect readout demonstrated by TATA-box binding protein. (Kim et al., 1993)

**Example of the complexity of DNA-protein interactions**

Previous structural studies have been undertaken to refine our understanding of protein-DNA interactions; most notable are the efforts of the Pabo lab on the Zif268 Zn-finger transcription factor. Extensive structural work with both wild-type and mutant versions of Zif268 has been done (Elrod-Erickson et al., 1998; Elrod-Erickson et al., 1996; Pavletich and Pabo, 1991). Optimized DNA substrates for the mutant Zif268 have even been determined and the structure of their complexes solved (Elrod-Erickson et al., 1998). At least 10 structures of Zif268-DNA complexes have been submitted to the pdb database (www.rcsb.org).

Zif268 consists of 3 $Cys_2His_2$ zinc fingers. Each of these fingers is about 30 amino acids long and consists of a two stranded anti-parallel β-sheet and an α-helix (the recognition helix) held together by a small hydrophobic core and a Zn ion coordinated between the secondary structure elements. Each finger recognizes 3 consecutive base-pairs in the target DNA, termed a subsite. To a rough approximation each zinc finger is essentially modular and recognizes its subsite independently of the other zinc finger modules. The recognition helix contributes up to 4 residues (numbered -1,2,3,6 in the helix) that are considered important for the specificity of each finger, these residues make contact to either the bases or the backbone on the major groove side of the DNA. Thus the majority of Zif268's specificity can be explained via direct readout. The architecture of these 3 zinc fingers is

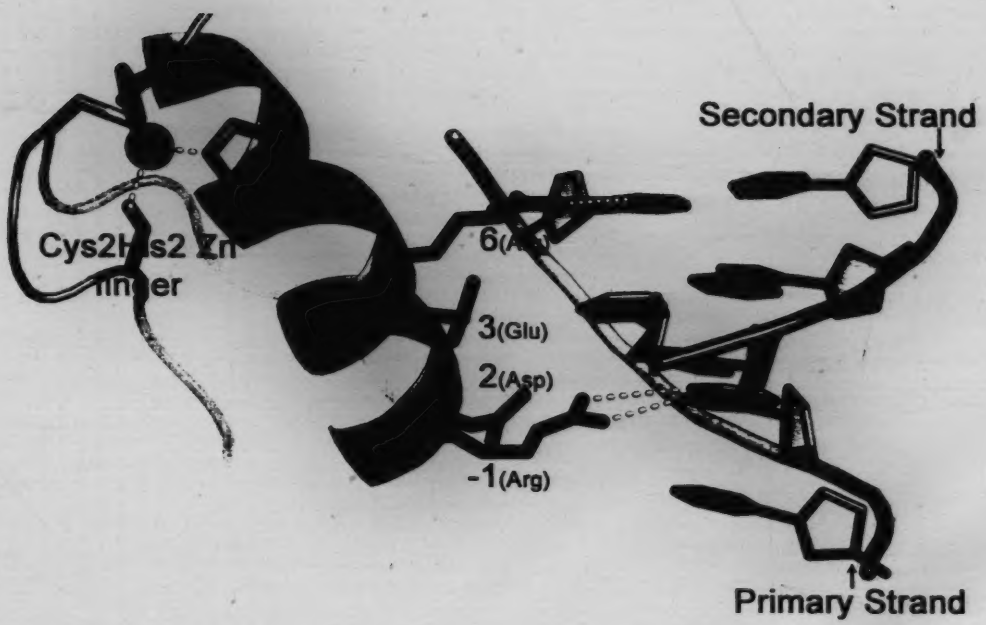


**Figure 1.3. Characteristic interactions of a Zif268-like zinc finger and its DNA subsite.** Residues positions are with respect to the start of the α helix. Not all contacts are made by every finger and some fingers have additional contacts as well.

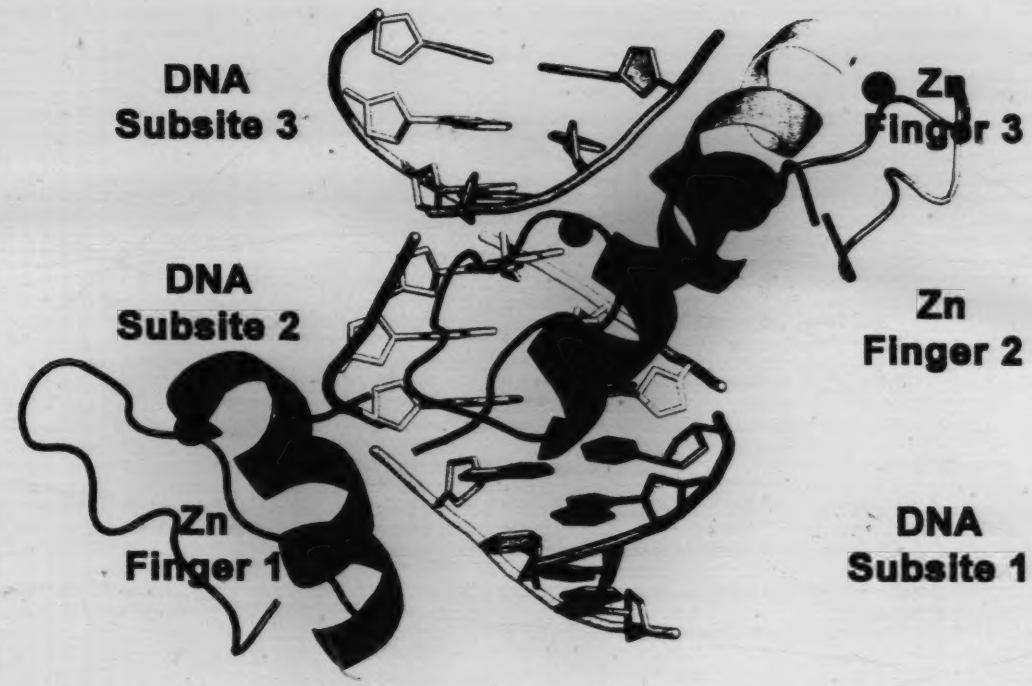


**Figure 1.4. The architecture of the 3 zinc fingers in the Zif268 transcription factor.** Finger 1 is colored blue with its DNA sub-site in orange. Finger 2 and its sub-site are teal and yellow respectively and finger 3 is green with its corresponding sub-site in light purple. Zinc ions are indicated by magenta spheres. Both the DNA and protein are continuous in the model, the gaps here are to emphasize the distinct and independent nature of each finger.

shown in figure 1.4. As a whole Zif268 recognizes a 9 base pair consensus as shown.

Because the positions in the helix that mediate recognition are conserved and each module was considered independent, the recognition of a DNA consensus appeared straightforward. Accordingly, the Pabo lab undertook a project to engineer variants of Zif268 that can recognize different DNA consensuses. They approached this project by using phage display selection experiments with Zif268 peptides randomized at the -1, 2, 3, 6 positions of zinc finger 1. Fingers 2 and 3 were left as wild type to serve as an anchor for the selection experiments. The results of this experiment were successful in engineering variants that bound their DNA sequences, however subsequent structural studies demonstrated that Zif268 does not utilize a simple "recognition code" where a particular side chain recognizes a corresponding base.

Often these changes had unpredicted results. For instance, the RADR peptide with its targeted GCA site binds tightly and specifically yet the only 2 direct base contacts are made at positions 2 and 6 of the helix. Ala(2) makes a van der Waals contact to a thymine methyl group 2 bases away from its subsite and Arg(6) makes a pair of hydrogen bonds to the guanine in the first position of its subsite. Surprisingly, Arg(-1) does not make a base contact but instead interacts with a phosphate while Asp(3) makes bifurcated hydrogen bonds to Arg(-1). Although this variant only makes 1 direct contact to its subsite it still binds with high affinity and specificity, indicating that indirect

recognition may play a role. How this recognition is accomplished was not discussed by the authors.

In addition, this RADR peptide binds tightly to the wild type site (GCG), the structure of which was also solved. In this case, both Arg(-1) and Asp(3) have alternate conformations. One is similar to the RADR - GCA site complex, in which the Arg(-1) contacts the backbone, while the other is similar to the wild type Zif268-DNA complex where Arg(-1) contacts the 3rd base (guanine) in its subsite (Figure 1.3). Comparison to the wild-type structure provides some explanation for these alternate conformations. In the wild-type structure, Asp(2) of the finger forms two hydrogen bonds to Arg(-1) stabilizing its conformation in a position that favors interaction with the guanine base. In the RADR variant this aspartate is replaced with an alanine and no longer fixes Arg(-1) in position to interact with the guanine base, resulting in the two conformations that are seen.

Even this seemingly simple system is full of complexities and the specificity at all positions in the consensus DNA cannot be fully explained. In addition to the detailed example given above, the authors observed atypical contacts to the DNA, contact between fingers and neighboring sub-sites and conformational flexibility of the zinc finger orientation relative to each other. All of these observations add complexity to understanding how transcription factors recognize their cognate DNA and throws a wrench into the prospect of predicting a proteins DNA preference *a priori*.

Although there are hundreds of protein-DNA complexes in the pdb database the vast majority are solved with optimal DNA substrates and many that are bound to suboptimal DNA lack similar structures for comparison. As a result there are few model systems available to dissect how proteins recognize their DNA targets. Many structures have been solved where a consensus DNA target is known and not explained by the complex structure. These discrepancies are commonly dismissed by invoking some vague version of indirect recognition or by the involvement of another unknown transcription factor. In this thesis, I attempt to uncover the more subtle modes of recognition used by Ndt80 to bind its consensus DNA.

**Ndt80 as a model system**

To understand how Ndt80 binds MSE DNA, we have used genetic and biochemical approaches to delineate the DNA binding domain of Ndt80 and have determined its structure bound to an MSE-containing oligonucleotide. We have refined this structure at a resolution limit of 1.4 Å, making it one of the highest resolution protein-DNA complexes solved to date. The structure reveals that Ndt80 is a member of the Ig-fold family of transcription factors, and recognizes DNA in a manner similar to other members of the family such as p53, NF-κB, NFAT, and STATs. However, unlike the other members of this family, Ndt80 can bind to its MSE substrate with high affinity as a monomer. This high affinity is likely due to the additional secondary structure elements present in Ndt80 that are not seen in other members of this family. The

integrity of the DNA binding domain is shown to be essential for transcriptional activation and normal progression through sporulation. The structure reveals that two conserved 5'-YpG-3' dinucleotide steps are recognized by arginine residues which simultaneously form hydrogen bonds to the guanine base and stack with the 5'-pyrimidine base, whereas the poly(A)-poly(T) portion of the MSE is recognized by minor groove interactions. A detailed analysis of the effects of mutations within the MSE on Ndt80-binding affinity reveals that both the 5'-YpG-3' dinucleotide steps and the poly(A)-poly(T) tract are critical specificity determinants for Ndt80 binding. These qualities of excellent resolution, monomeric high affinity binding, and multiple modes of protein-DNA interactions make Ndt80 an excellent candidate for use as a model system, to further our understanding of protein-DNA interactions.

| | |
|---|---|
| Ndt80 | 5'-GNCACAAAA-3' |
| | 3'-CNGTGTTTT-5' |
| p53 | 5'-GGGCAAGT-3' |
| | 3'-CCCGTTCA-5' |
| NFAT | 5'-GGAAAA-3' |
| | 3'-CCTTTT-5' |
| NF-κB (p50) | 5'-GGGGAA-3' |
| | 3'-CCCCTT-5' |
| NF-κB (p65) | 5'-GGAAT-3' |
| | 3'-CCTTA-5' |
| STAT | 5'-GGGAA-3' |
| | 3'-CCCTT-5' |

**Figure 1.5 Similar DNA substrates of other Ig fold transcription factors.** These listed transcription factors have a 5'-GC rich region and a 3'-AT rich portion that are recognized in similar ways.

**Research overview**

Early work with Ndt80 focused on genetic studies that were used to identify this protein and establish its role as a middle sporulation transcription factor. We began work on Ndt80 primarily in a structural capacity to determine how this protein recognizes the MSE. The structural work was aimed at answering several questions: 1. What is the core fold of Ndt80? 2. What are the key protein DNA contacts that mediate MSE recognition? 3. Can this structural information be used to postulate the relative affinity that results from variations of the MSE? This thesis describes my work towards solving the structures of the core domain of Ndt80 and the complex of the DNA-binding domain of Ndt80 bound to an MSE. It describes the insight we gained from the solution of these structures and subsequent structures of Ndt80 complexed with variant and mutant MSE that we solved fully address our questions more fully. In all I solved 12 structures that illustrate not only how Ndt80 recognizes its target but allowed us to propose mechanisms of recognition that are employed by other transcription factor families as well.

Chapter 2 describes the genetic and biochemical work to define the DNA binding domain of Ndt80 in addition to a core domain within the DNA binding region. It continues with the work that resulted in structures of the core domain and of the DNA-binding domain in complex with an MSE-containing DNA.

Chapter 3 continues the analysis of the DNA binding determinants by focusing on the recognition of YpG steps by arginine sidechains. I show

evidence that this mode of interaction is important not only for Ndt80-MSE recognition but also for numerous other protein - DNA complexes.

Despite the efforts described in the previous chapters we still did not fully understand how Ndt80 bound to the MSE. Issues such as preference for poly A sequence rather than mixed A-T sequences at the 3' end and selectivity at the ambiguous positions still plagued us. In Chapter 4, I describe my efforts to address these concerns structurally by solving the structures of Ndt80 bound to 10 different variant and mutant MSE sequences. The resulting high resolution complexes provided enormous detail and possible explanations for these questions.

In the final chapter (Chapter 5), I discuss the comparison of structure and function of Ndt80 to other Ig-fold transcription factors. Following this, I address the role Ndt80 plays in meiosis and rationalize the *in vivo* expression profile of Ndt80 controlled middle genes with the help of computer models. Next, I discuss the contribution these structures have made in understanding DNA-protein recognition, not only in this system but in the general case. Finally, I propose some future work that I feel would be beneficial both in future modeling of DNA-protein structures and understanding the interactions between such molecules.

## REFERENCES


Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., and Harrison, S. C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science *242*, 899-907.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994). Molecular Biology of the Cell, Third edn (New York: Garland Publishing).

Anderson, J. E., Ptashne, M., and Harrison, S. C. (1987). Structure of the repressor-operator complex of bacteriophage 434. Nature *326*, 846-852.

Bell, G. (1982). The Masterpiece of Nature (Berkeley: Univ. California Press).

Cavalli, P., Bosi, A., and Bassi, D. (2003). Down's syndrome. Lancet *362*, 81.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Chu, S., and Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. Mol Cell *1*, 685-696.

Elrod-Erickson, M., Benson, T. E., and Pabo, C. O. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. Structure *6*, 451-464.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure *4*, 1171-1180.

Goddard, M. R., Godfray, H. C., and Burt, A. (2005). Sex increases the efficacy of natural selection in experimental yeast populations. Nature *434*, 636-640.

Hepworth, S. R., Ebisuzaki, L. K., and Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in Saccharomyces cerevisiae. Mol Cell Biol *15*, 3934-3944.

Kassir, Y., Adir, N., Boger-Nadjar, E., Raviv, N. G., Rubin-Bejerano, I., Sagee, S., and Shenhar, G. (2003). Transcriptional regulation of meiosis in budding yeast. Int Rev Cytol *224*, 111-171.

Kassir, Y., Granot, D., and Simchen, G. (1988). IME1, a positive regulator gene of meiosis in S. cerevisiae. Cell *52*, 853-862.

Kim, Y., Geiger, J. H., Hahn, S., and Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512-520.

Kupiec, M., Byers, B., Esposito, R. E., and Mitchell, A. P. (1997). Meiosis and Sporulation in *Saccharomyces cerevisiae*, In The Molecular and Cellular Biology of Yeast, J. Pringle, J. Broach, and E. Jones, eds. (Cold Spring Harbor, Ney York: Cold Spring Harbor Laboratory Press), pp. 889-1036.

Lejeune, J., Gautier, M., and Turpin, R. (1959). Study of somatic chromosomes from 9 mongoloid children. Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences *248*, 1721-1722.

Mitchell, A. P. (1994). Control of meiotic gene expression in Saccharomyces cerevisiae. Microbiol Rev *58*, 56-70.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321-329.

Ozsarac, N., Straffon, M. J., Dalton, H. E., and Dawes, I. W. (1997). Regulation of gene expression during meiosis in Saccharomyces cerevisiae: SPR3 is controlled by both ABFI and a new sporulation control element. Mol Cell Biol *17*, 1152-1159.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

Pierce, M., Benjamin, K. R., Montano, S. P., Georgiadis, M. M., Winter, E., and Vershon, A. K. (2003). Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol *23*, 4814-4825.

Pierce, M., Wagner, M., Xie, J., Gailus-Durner, V., Six, J., Vershon, A. K., and Winter, E. (1998). Transcriptional regulation of the SMK1 mitogen-activated protein kinase gene during meiotic development in Saccharomyces cerevisiae. Mol Cell Biol *18*, 5970-5980.

Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A *73*, 804-808.

Shuster, E. O., and Byers, B. (1989). Pachytene arrest and other meiotic effects of the start mutations in Saccharomyces cerevisiae. Genetics *123*, 29-43.

Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D., and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci U S A *102*, 1998-2003.

Watson, J. D., and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Xie, J., Pierce, M., Gailus-Durner, V., Wagner, M., Winter, E., and Vershon, A. K. (1999). Sum1 and Hst1 repress middle sporulation-specific gene expression during mitosis in Saccharomyces cerevisiae. Embo J *18*, 6448-6454.

Xu, L., Ajimura, M., Padmore, R., Klein, C., and Kleckner, N. (1995). NDT80, a meiosis-specific gene required for exit from pachytene in Saccharomyces cerevisiae. Mol Cell Biol *15*, 6572-6581.

# Chapter 2:

# The crystal structures of the core domain of Ndt80 and the DNA - binding domain of Ndt80 bound to DNA

**SUMMARY**

Progression through the middle phase of sporulation in *Saccharomyces cerevisiae* requires sporulation specific transcription factor, Ndt80. Ndt80 binds a specific DNA sequence, the middle sporulation element (MSE), and activates the transcription of ~ 150 middle genes. Using a combination of deletion analysis and limited proteolysis we have defined a structurally ordered core DNA-binding domain of Ndt80. Further biochemical studies identified a region that is required, in addition to the core region, for full affinity and specificity of Ndt80. We have crystallized and solved the structure of both the core DNA-binding domain and the full DNA-binding domain in complex with an MSE containing DNA. The structures reveal that Ndt80 is a member of the Ig-fold family of transcription factors. The structure of the DNA-bound form was refined at a resolution of 1.4 Å, one of the highest resolution protein-DNA complexes solved, thus giving unprecedented detail for analysis. For example, the complex reveals an unexpected mode of recognition whereby a 5'-pyrimidine-guanine-3' dinucleotide step is recognized by a single arginine residue that simultaneously hydrogen bonds to the 3' guanine base while forming stacking/van der Waals interactions with the 5' pyrimidine. Analysis of the DNA-binding affinity of MSE mutants demonstrates the importance of these interactions and highlights other important regions of the MSE such as the poly-A region. The majority of this work was originally published in EMBO Journal (Lamoureux et al., 2002)

## INTRODUCTION

The transcriptional cascade of sporulation in yeast requires the precise timing of gene expression to coordinate the simultaneous processes of spore morphogenesis and meiosis. We wanted to understand how Ndt80 recognizes the appropriate promoter elements in the middle genes it induces. Although 1/3 or approximately 2000 genes in the yeast genome contain an MSE site in their promoter region, it is estimated only ~ 150 are actually under control of Ndt80(Chu et al., 1998; Jolly et al., 2005). Although many other factors (chromatin state, repressors, etc.) contribute to this discrepancy, understanding the protein-DNA recognition at a molecular level could help to limit the substrate definition. Ndt80 has very little sequence homology with any organisms outside of the fungi kingdom, and certainly no homology among proteins with known structures. With no help from available databases we undertook experiments to delineate the DNA binding domain of Ndt80 and subsequently solve the structure.

## EXPERIMENTAL PROCEDURES

### Cloning and Vector Construction

Template DNA for Ndt80 was from a genomic DNA preparation from the SK1 strain of *S. cerevisiae*. Polymerase chain reaction (PCR) was used to amplify the template using the chemically synthesized oligonucleotides. The constructs used in this chapter are summarized below in the format; gene construct (named and numbered as the resulting protein residues - not nucleotides), 5' oligonucleotide (N-terminus of the resulting protein), 3' oligonucleotide (C-terminus of the resulting protein);

**Ndt80(1-627)**, 5'-cca gat cta tga atg aaa tgg aaa ac-3', 5'-ggt gtc gac tta ata ctt ata gaa act atc-3'; **Ndt80(1-349)**, 5'-caa gat cta tga atg aaa tgg aaa ac-3', 5'-ccg gtg gtc gac tta ttg tga gga att gac act cga-3'; **Ndt80(350-627)**, 5'-gcg tgg aga tct aac agc aca aaa aga aaa atg-3', 5'-ggt gtc gac tta ata ctt ata gaa act atc-3'; **Ndt80(185-530)**, 5'-gcg tgg aga tct cct tca gta tgt ccg ttg gtg-3', 5'-ccg gtg gtc gac tta tgt ttt gca ttc aga acg tga-3'; **Ndt80(185-627)**, 5'-gcg tgg aga tct cct tca gta tgt ccg ttg gtg-3', 5'-ggt gtc gac tta ata ctt ata gaa act atc-3'; **Ndt80(1-530)**, 5'-caa gat cta tga atg aaa tgg aaa ac-3', 5'-ccg gtg gtc gac tta tgt ttt gca ttc aga acg tga-3'; **Ndt80(1-340)**, 5'-caa gat cta tga atg aaa tgg aaa ac-3', 5'-ccg gtc gac tta tca tct cac agt tat tcg-3'

### Protein Expression and Purification

Full-length Ndt80 was expressed in *E. coli* as a fusion with maltose binding protein (MBP) using the expression plasmid, pMAL-NDT80. A 2.4 Kb

genomic DNA fragment encoding the *NDT80* gene was used as a template to produce the *NDT80* open reading frame for pMAL-NDT80. The *NDT80* open reading frame was amplified with Vent DNA polymerase (New England Biolabs) and ligated between the BamHI and SalI sites in the expression vector pMAL-C2 (New England Biolabs). MBP-Ndt80 fusion protein was expressed and purified by maltose affinity chromatography as previously described (Chu and Herskowitz, 1998).

Fragments of Ndt80 were expressed in *E. coli* as fusion proteins with GST. For GST-Ndt80(1-340), the region of *NDT80* encoding residues 1–340 was amplified by PCR and cloned into pGEX-6P1 (Amersham BioSciences), between the BamHI and XhoI restriction sites. For GST-Ndt80(59-340), the region encoding residues 59-340 was cloned into pGEX-KG. Both proteins were expressed in *E. coli* BL21 with a schemetic for Ndt80(1-340) shown in Figure 2.1 . Fresh 2xYTA media was inoculated with 50 mL/L of saturated overnight culture and grown at 30°C to an $OD_{600} \approx 1$. The culture was induced with a final concentration of 0.25 mM IPTG and grown an additional 6 hrs at 30°C. The cells were harvested and flash frozen in liquid nitrogen. Cells were then resuspended in PBS (pH 7.3), 1 mM EDTA, 5 mM DTT, pepstatin, leupeptin, benzamidine, PMSF, 0.5 mg/mL lysozyme and gently mixed on ice for 30 min., followed by sonication. The lysate was cleared by centrifugation and the supernatant was then mixed at room temperature with glutathione sepharose 4B beads for 1 hr. This slurry was poured into a column and washed with 50 bed volumes of PBS containing 1 M NaCl, followed by 50 bed

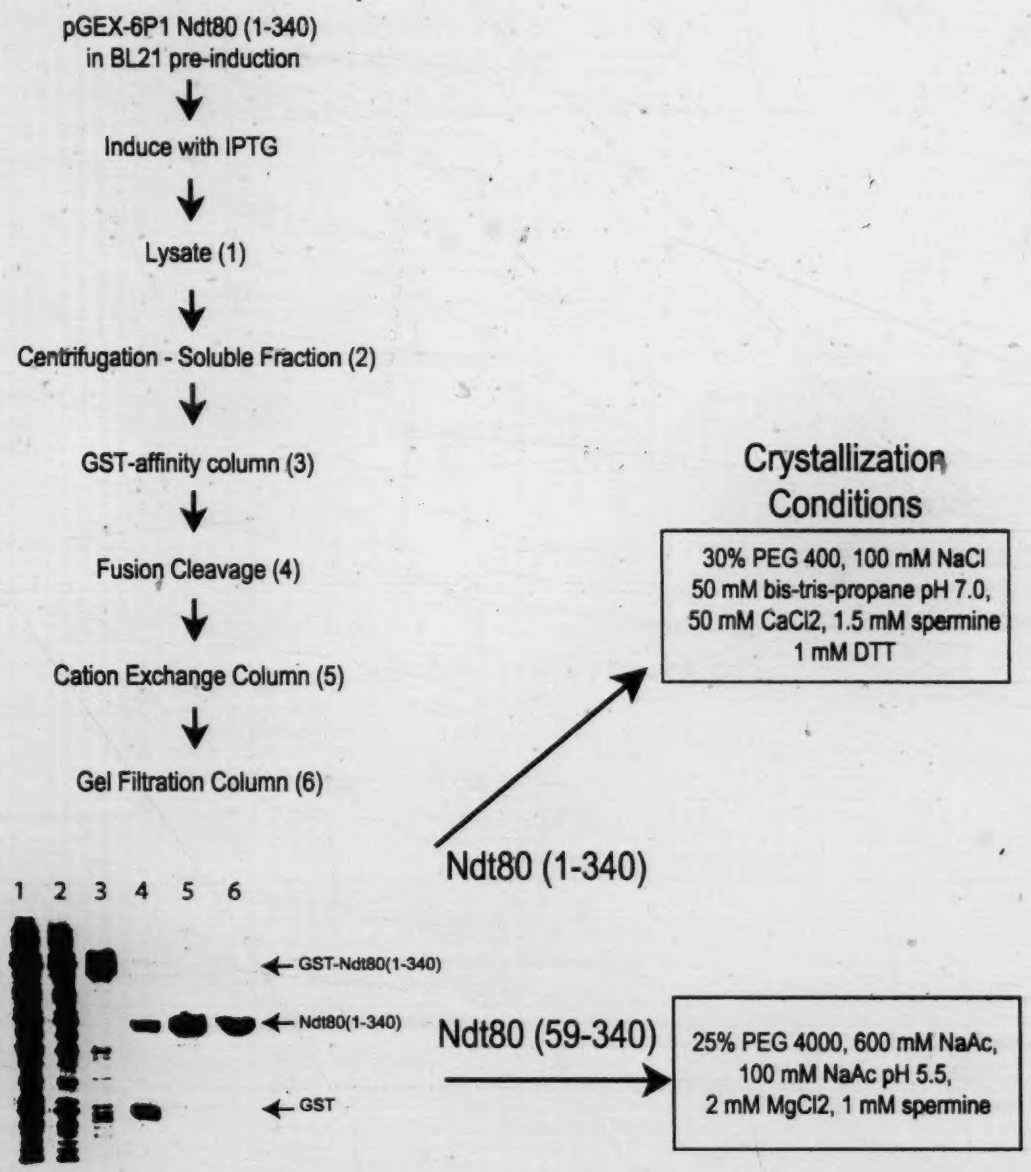


**Figure 2.1 Expression, purification and crystallization of Ndt80(59-340) and Ndt80(1-340).** Ndt80 (1-340) is focused on in this figure although there are only minor changes to the procedure for Ndt80 (59-340). The GST-fusion protein is purified by affinity chromatography, followed by cleavage of the affinity tag and two subsequent chromatographic steps. The final chromatographic step is a size exclusion column which primarily serves as a buffer exchange step as there is rarely any additional purification at this stage.

volumes of PBS. Fusion protein was eluted with PBS (pH 8.0) containing 20 mM glutathione. Ndt80(1-340) was liberated from GST by cleavage with PreScission protease (Amersham BioSciences) at 4°C overnight, while Ndt80(59-340) was liberated from GST-Ndt80(59-340) by thrombin cleavage. Protein was further purified by cation exchange (SP sepharose FF) and gel filtration (Superdex 75) chromatography. The resulting protein was concentrated to 20 mg/mL in a buffer containing; 10 mM Tris (pH 7.0), 100 mM NaCl, 1 mM DTT.

**Electrophoretic mobility shift assays (EMSA)**

All of the EMSA were performed on chemically synthesized blunt end 20-mer DNA using the MSE or MSE variant as shown, with flanking sequences matching the *SPS4* MSE flanking sequences. Duplex DNA was radiolabelled using T4 kinase and $\gamma$-$^{32}$P ATP. All the binding reactions contained 75 mM KCl, 10 mM HEPES (pH 7.9), 11 mM $MgCl_2$, 50 $\mu$M $ZnSO_4$, 10 % glycerol, 1 mM DTT, 0.25 mM EDTA, 0.025% bromophenol blue and 0.25 nM radiolabelled MSE DNA (Hepworth et al., 1998). All protein dilutions were made using 10 mM Tris (pH 7.0), 100 mM NaCl, 1 mM DTT, 0.1 mg/mL BSA. Purified Ndt80(1-340) was the final component added to the binding reaction, which was subsequently equilibrated for 15 min. at room temperature. The reaction was then loaded onto a 8% polyacrylamide gel pre-run at 100 V for 30 min. in 0.5 X TBE and run for another 2 hrs at 150 V. Bands were then visualized and quantified using phosphor imaging plates and

the program ImageQuant (Molecular Dynamics). The competition experiments were performed under the same conditions as above, but with the $^{32}P$ DNA and protein given a 5 min. pre-incubation before competitor DNA was added. After addition of competitor, the reaction was incubated 15 min. at 20°C and loaded onto the gel. The initial EMSA used in the deletion analysis were performed with induced, soluble cell lysates from *E. coli* transformed with the various deletion mutants in the presence of 10-fold weight excess of poly(dI-dC).

**Proteolytic mapping**

Flexible regions in Ndt80(1-349) were characterized by limited proteolytic mapping (Figure 2.2). Purified GST-Ndt80(1-349) was subjected to digestion with trypsin or thrombin over a 24 hr time course either alone or in the presence of a 5-fold molar excess of MSE DNA. Stable fragments were subsequently purified by cation exchange (SP sepharose FF) and gel filtration (Superdex 75) chromatography and their masses were determined by MALDI-TOF mass spectrometry. This analysis indicated that Ndt80(59-340) is the major trypsin/thrombin resistant Ndt80 fragment in the absence of DNA, while residues 1-340 is protected when bound to the MSE.

**DNA purification**

Synthetic DNA oligonucleotides were purified on a Source 15Q column under denaturing conditions (10 mM NaOH), desalted with a C18 cartridge in

a volatile buffer, lyophilized and resuspended in 5 mM Tris pH 7.0. DNA solutions were quantified by absorbance at 260 nm. Duplex DNA was annealed by heating to 80°C and slow cooling to room temperature in 5 mM Tris (pH 7.0), 100 mM NaCl at a final concentration of 1 mM DNA duplex (Figure 2.3).

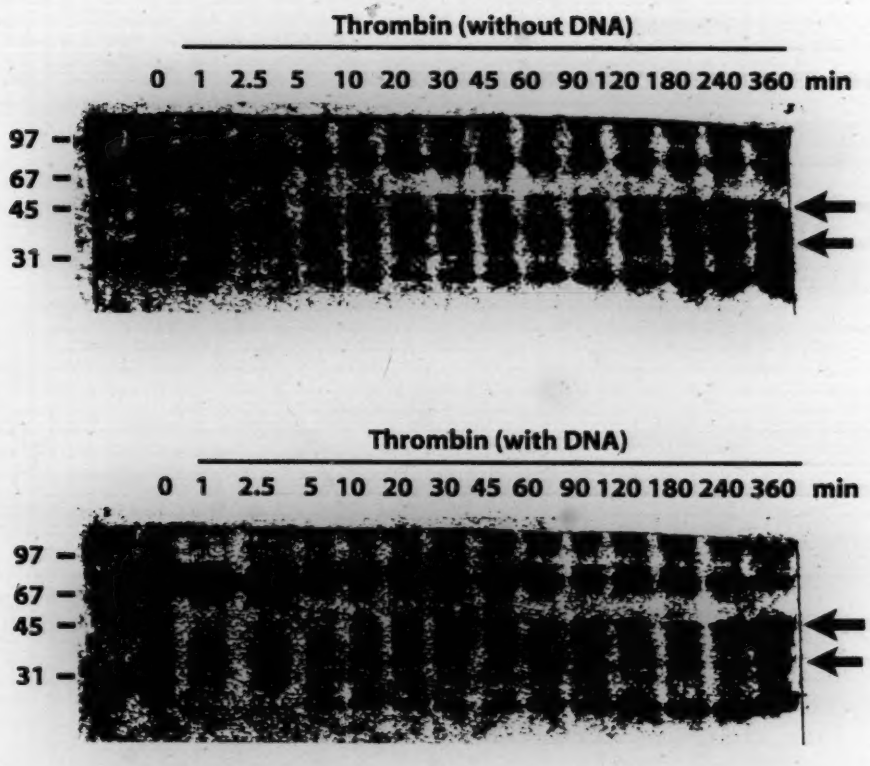


**Figure. 2.2 Limited proteolysis of Ndt80 shows a region on the N-terminus that is stabilized upon the addition of MSE containing DNA.** Purified Ndt80 (1-349) amino terminally fused to GST was digested with thrombin in the presence and absence of DNA for the indicated times. The gray arrow indicates the position of Ndt80 (59-340) and black arrow of Ndt80 (1-340). The fastest migrating major band is GST.

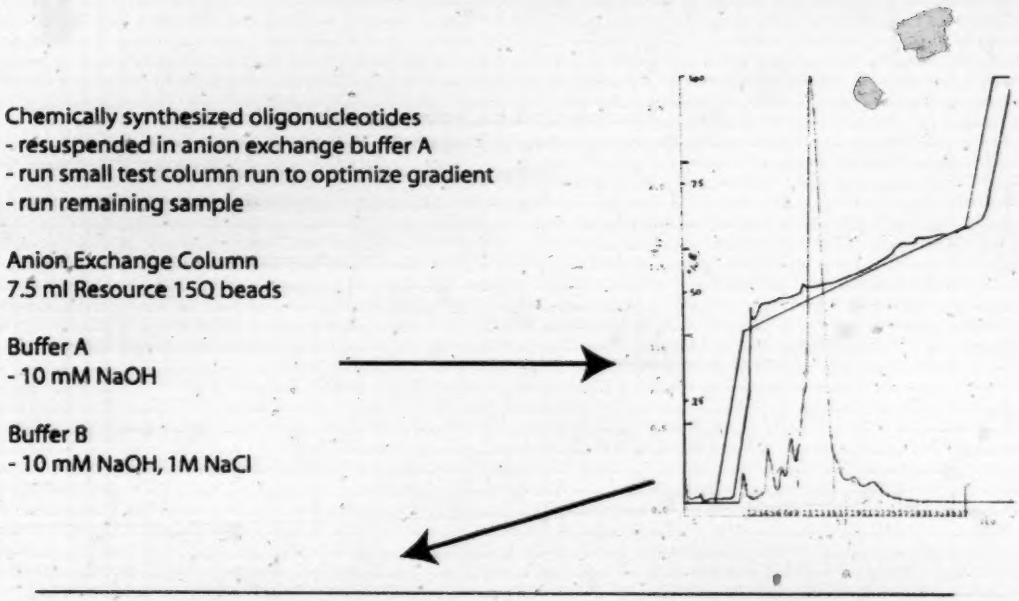
Chemically synthesized oligonucleotides
- resuspended in anion exchange buffer A
- run small test column run to optimize gradient
- run remaining sample
Anion Exchange Column
7.5 ml Resource 15Q beads
Buffer A
- 10 mM NaOH
Buffer B
- 10 mM NaOH, 1M NaCl
Pooled fractions of the central peak
and de-salt using Sep-Pak (C18 cartriges)
Desalting Procedure
Solutions
1. Acetonitrile
2. Elution Buffer (E) 30% acetonitrile,
0.1 M TEAB (triethylammonium bicarbonate)
3. Wash Buffer (W) 0.25mM TEAB
Pre-equilibrate, Load, Wash and Elute
1. 1 x 10 mL acetonitrile
2. 1 x 10 mL Elution Buffer
3. 2 x 10 mL Wash Buffer
4. Load Sample at 1-2 drops/second
5. Wash 2 x 10 mL wash buffer
6. Elute with 5 mL elution buffer
Lyophilize O/N, resuspend in H2O,
SpeedVac, resuspend in 10 mM tris pH 7.0
Each oligo was purified, desalted, and
quantitated individually by absorbance
at 260 nm. Then complimentary oligos
were annealed in equimolar amounts
by heating to 80 degreec C and slow
to room temperature. Binding to
Ndt80(1-340) was tested with native
mini gels
EtBr stained
Coomassie Stained
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9

**Figure 2.3 DNA purification scheme.** Purification, desalting, quantification, annealing and verification of binding. Top image is a sample chromatograph of an anion exchange column. Bottom gel images are of the same gel; one stained with EtBr the other with Coomassie Blue. The gel contains (1) protein alone (2) duplex DNA alone and 7 DNA Ndt80(1-340) complexes ranging in size from 21-11 basepairs (3-9). Note that the last lane containing an 11mer duplex binds poorly and may be the minimal size DNA suitable for crystallographic study.

**Crystallization and data collection**

Ndt80(1-340)-MSE complexes were prepared to a protein concentration of 10 mg/mL at a molar ratio of protein:DNA of 1:1. Crystals were grown using the hanging drop vapor diffusion method at room temperature (20°C) in conjunction with streak seeding (Figure 2.1, 2.5). 2 μL of reservoir solution (30% PEG 400, 50 mM bis-tris-propane pH 7.0, 100 mM NaCl, 50 mM $CaCl_2$, 1.5 mM spermine, 1 mM DTT) and 2 μL Ndt80-MSE complex were mixed and immediately streaked with a hair dipped in streak solution and rinsed twice in the reservoir. Streak solution was prepared by a ~50-fold dilution of a drop that contained a shower of small crystals. Crystals grew to a maximal size of 400 μm in about 1 week and were harvested and frozen in reservoir solution. Data were collected at beamline 9-2 at Stanford Syncrotron Radiation Laboratory on an ADSC Q4 CCD. The crystals belong to spacegroup $C222_1$ (a=70.13 Å, b = 78.81 Å, c = 161.39 Å) with one molecule in the asymmetric unit. Data were processed with DENZO and SCALEPACK (Otwinowski and Minor, 1997)(Table 2.1).

Ndt80(59-340) crystals were grown using the hanging drop method at 20°C from 4 mg/mL protein solutions (which also contained a 20-mer DNA duplex at a 1:1 molar ratio), equilibrated against a reservoir solution of 600 mM sodium acetate, 2 mM $MgCl_2$, 1 mM spermine, 25% PEG 4000, 100 mM sodium acetate pH 5.5 (Figure 2.1, 2.4). Equal amounts of protein and reservoir solutions were mixed and crystals grew to a maximal size of 300 μm

A

B

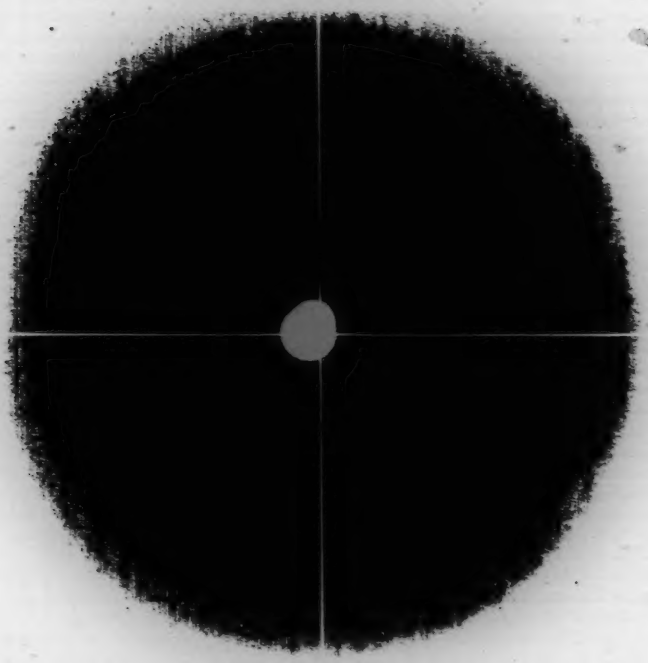

**Figure 2.4 Crystals and diffraction from Ndt80(59-340). (A)** Crystals of Ndt80(59-340) the panel on the left is representative of crystals used to obtain the data set and are approximately 100 microns across (B) Sample diffraction pattern of the Se-methionine crystals at the high energy remote wavelength. The edge of the plate is approximately 2.3 Å.

A

B

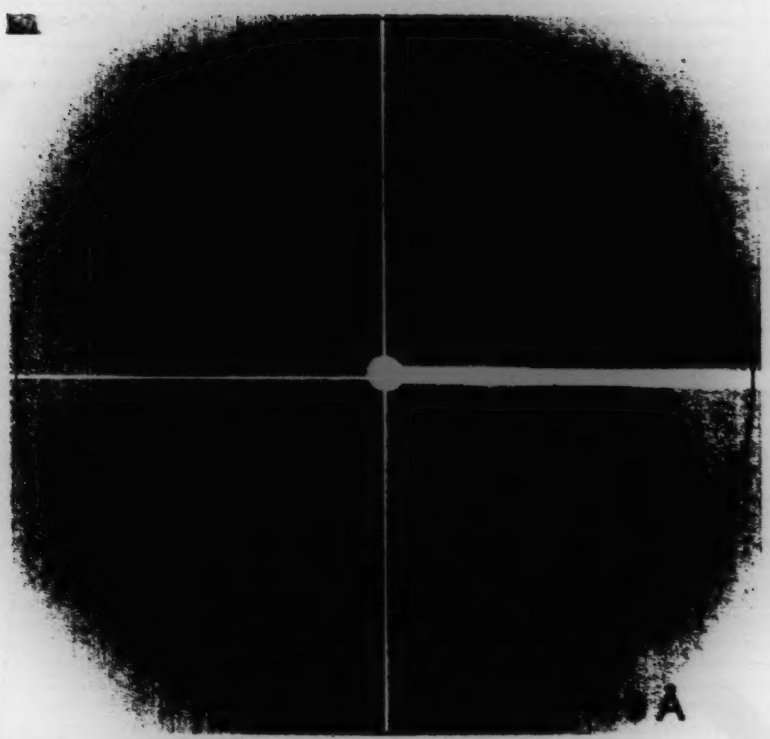


**Figure 2.5 Ndt80 (1-340) MSE complex. A.** Ndt80(1-340) and 14mer MSE containing DNA complex crystals. The scale applies only to the central picture, the peripheral pictures are enlarged approximately two fold relative to the middle. **B.** Sample diffraction pattern of the native crystals, this is an image from the low resolution pass, a high resolution pass was also done with longer exposures.

**Table 2.1. Summary of X-ray experiments**

| **Crystal data** | **Ndt80(1-340) + MSE DNA** | | | | **Ndt80 (59-340)** | | |
|---|---|---|---|---|---|---|---|
| a (Å) | 70.13 | | | | 35.88 | | |
| b (Å) | 78.81 | | | | 41.93 | | |
| c (Å) | 161.39 | | | | 163.91 | | |
| Space group | $C222_1$ | | | | $P2_12_12_1$ | | |
| **Data Collection** | Native | MAD $\lambda_1$ | MAD $\lambda_2$ | MAD $\lambda_3$ | MAD $\lambda_1$ | MAD $\lambda_2$ | MAD $\lambda_3$ |
| Wavelength (Å) | 0.98008 | 0.97936 | 0.97922 | 0.93218 | 0.9793 | 0.9789 | 0.9563 |
| Resolution | 100-1.4 | 100-1.7 | 100-1.7 | 100-1.7 | 40-2.2 | 40-2.2 | 40-2.2 |
| Reflections | | | | | | | |
| total | 3338673 | 218954 | 219186 | 220867 | 141688 | 141567 | 144249 |
| unique | 88384 | 49236 | 49277 | 49065 | 12360 | 12286 | 12292 |
| Completeness | 99.8(97.8) | 98.6(87.3) | 98.4 (82.8) | 98.1(90.1) | 99.1(100) | 97.8(100) | 98.1(100) |
| I/σ | 35 (3.0) | 18.5 (3.5) | 18.1 (3.1) | 18.1 (3.9) | 19.9 (7.3) | 17.5 (5.4) | 20.9 (7.9) |
| Rsym | 5.1(49.3) | 4.5 (23.8) | 5.3 (29.7) | 4.5 (30.4) | 8.6 (25.4) | 9.9 (32.7) | 7.6 (22.1) |
| Redundancy | 38 (5.2) | 4.5 (2.1) | 4.5 (2.1) | 4.5 (2.6) | 11.5 (5.7) | 10.5 (5.2) | 11.7 (5.4) |
| **Refinement** | **Ndt80(1-340) + MSE DNA** | | | | **Ndt80 (59-340)** | | |
| Rcryst/Rfree | 19.4/20.6 | | | | 22.6/26.8 | | |
| RMSD bonds (Å) | 0.011 | | | | 0.006 | | |
| RMSD angles (°) | 1.52 | | | | 1.28 | | |
| Ramachandran plot (% residues in region) | | | | | | | |
| favored | 90.6 | | | | 85.3 | | |
| allowed | 8.2 | | | | 12.4 | | |
| generously allowed | 0.8 | | | | 1.8 | | |
| disallowed | 0.4 | | | | 0.5 | | |

Values in parentheses are statistics for the highest resolution shell, (1.42-1.40Å for native and 1.73-1.70Å for MAD data of Ndt80(1-340) and 2.25-2.20 Å for all data sets of Ndt80(59-340)).

$R_{sym} = 100\Sigma_h\Sigma_i| I_i(h)-<I(h)>| / \Sigma_h\Sigma_i I(h)$, for the intensity of i observations of reflection h.

$R_{cryst} = \Sigma_h|F_O(h) - F_C(h)| / \Sigma_h | F_O(h)|$, where $F_O(h)$ and $F_C(h)$ are observed and calculated structure factors.

$R_{free}$ calculated with 5% of all reflections excluded from refinement stages using high-resolution data.

Note: Ndt80(59-340) was refined against the MAD λ3 (high energy remote) data set as no high quality native data set was collected.

in 2 weeks. Cryoconditions are equivalent to the reservoir solution plus 15% glycerol. Data were collected at the Advanced Photon Source, BioCARS beamline 14–BM-D on an ADSC Q4 CCD and processed with DENZO and SCALEPACK (Table 2.1)

**Structure determination and refinement**

The DNA-bound and unbound Ndt80 structures were each independently solved using a three-wavelength seleno-methionine MAD experiment (Table 2.1). Selenium sites were located using SOLVE (Terwilliger and Berendzen, 1999) and phases were improved using both RESOLVE (Terwilliger, 2000) and DM (Cowtan, 1994). The unbound structure was built manually using O (Jones et al., 1991), and refined against λ3 of the MAD data set with iterative cycles of CNS refinement (Brunger et al., 1998) and manual rebuilding.

The initial model of Ndt80 bound to DNA was created using ARP/wARP (Perrakis et al., 1999) with further manual building done with O. Automated refinement was carried out with CNS and REFMAC (Collaborative Computational Project, 1994) and protein geometry was analyzed with PROCHECK (Laskowski et al., 1993). The final Ndt80-MSE model contains Ndt80 residues 33-139, 146-286, 294-335 and all of the DNA bases (although poor electron density is observed for the overhanging nucleotides T(–3) and A12'). The final Ndt80-MSE model contains 345 water molecules, with very good electron denisty as shown in Figure 2.6. The atomic coordinates of the

Ndt80-MSE complex, and the free Ndt80 structure, have been deposited in the Protein Data Bank (PDB ID: 1MNN (Ndt80(1-340)-MSE), 1MN4 (Ndt80(59-340)).

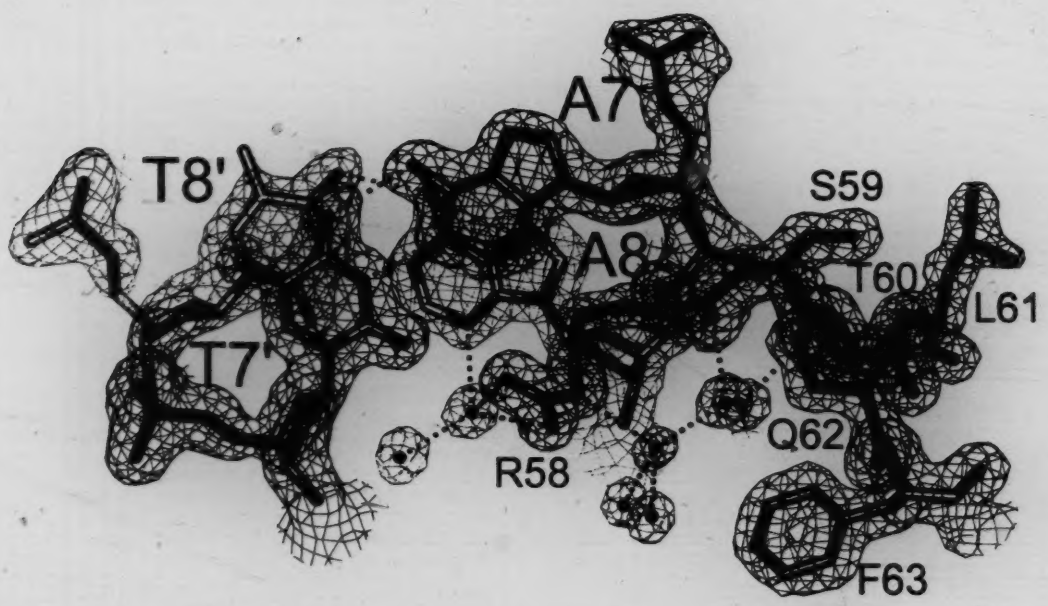


**Figure 2.6 Sample electron density.** $2|F_o|-|F_c|$ electron density map contoured at 1.5 σ of Ndt80(1-340)-MSE complex. The DNA map is contoured in blue, the protein in purple and the water molecules in green.

## RESULTS AND DISCUSSION

### Elucidation of the Ndt80 DNA binding domain

Ndt80 bears little detectable sequence similarity to other proteins that might shed light on its structure or function. Using deletion analysis, we isolated an N-terminal DNA-binding fragment of Ndt80 (Ndt80(1-349)) and, using limited proteolysis, we were able to show that residues 59–340 adopt a proteolytically resistant conformation in the absence of DNA, while the N-terminal residues 1-58 only become protected against proteolysis when bound to DNA (Figure 2.2). We next compared the abilities of full length Ndt80, Ndt80(1-340), and Ndt80(59-340) to specifically bind to an MSE-containing DNA in an electrophoretic mobility shift assay (EMSA). The results of this experiment (Figure 2.7) show that while all three proteins are capable of binding the MSE, only full length Ndt80 and Ndt80(1-340) form a complex that is resistant to challenge with a non-specific competitor DNA (poly(dI-dC)). Thus, residues 1-58 of Ndt80 are essential for sequence-specific recognition of the MSE and we conclude that Ndt80(1-340) has a minimal DNA binding domain that contains all the protein elements necessary for specific MSE recognition.

### Binding affinity of Ndt80 for mutant MSEs

Ndt80 activates over 150 distinct genes to allow the progression through meiotic prophase I (Chu et al., 1998). A comparison of the upstream regulatory sequences of a number of genes that are activated by Ndt80

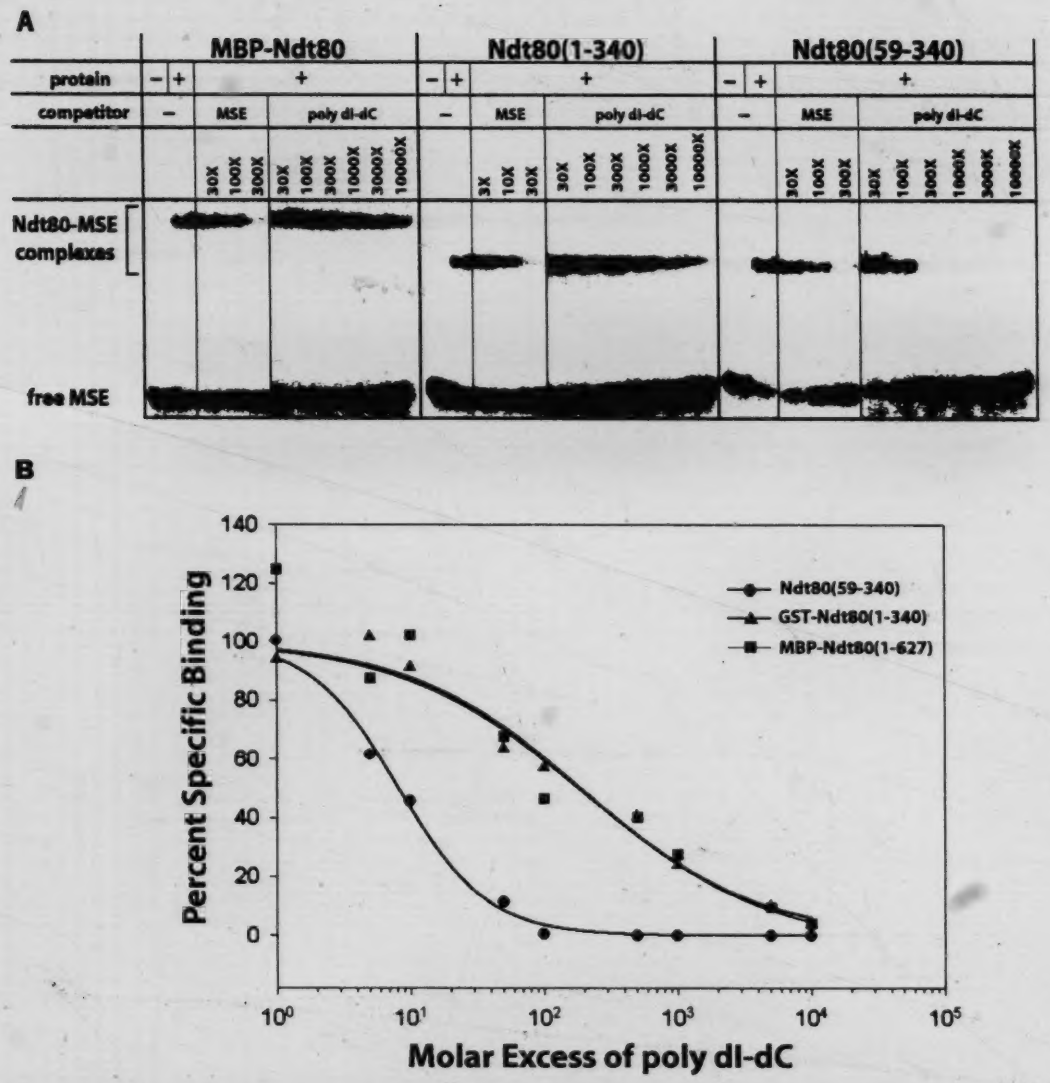


**Figure 2.7 Defining the minimal DNA binding domain of Ndt80.** (A) Purified recombinant MBP-Ndt80, Ndt80(1-340), or Ndt80(59-340), was bound to $^{32}$P-labelled MSE DNA and challenged with either the unlabelled MSE DNA as a specific competitor, or poly (dI-dC) as non-specific competitor at a variety of molar ratios of competitor to $^{32}$P-labelled MSE DNA as indicated. Note that an ~10-fold lower concentration of Ndt80(1-340) than MBP-Ndt80 was used to bind $^{32}$P-labelled MSE DNA. As a result, ~10-fold less of either cold DNA is required to compete the Ndt80(1-340)-MSE complex, compared to the MBP-Ndt80-MSE complex. (B) Graphical representation of specific binding in panel A, demonstrating that Ndt80 1-340 has the same binding characteristics as full length Ndt80.

revealed a 9 base pair MSE consensus sequence: 5'-gNCRCAAA(A/T)-3' (where "g" represents a non-conserved guanine nucleotide, "R" represents a purine nucleotide, and "N" represents a non-conserved position) (Chu et al., 1998; Hepworth et al., 1995; Ozsarac et al., 1997). Here we label the bases 1 through 9 in this strand of the MSE in the 5'-3' direction; the complementary bases in the opposite strand are given the same numbers but are distinguished with a "prime" (Figure 2.9). To understand the relative importance of individual base pairs within the MSE to Ndt80-binding affinity, we created a set of MSE-containing DNA oligonucleotides in which each position within the MSE was individually replaced. The DNAs were based on the MSE from the *SPS4* promoter, which exactly matches the MSE consensus, and has previously been shown to bind Ndt80 and activate transcription (Chu and Herskowitz, 1998). Ndt80 binding to the mutated DNAs was assayed by electrophoretic gel mobility shift experiments. A graph summarizing the relative affects of these mutations on Ndt80-binding affinity are shown in Figure 2.8. The results show that while the substitution of any of the consensus MSE base pairs leads to a reduction in Ndt80-binding affinity, by far the most dramatic substitution was the replacement of the C5-G5' pair with G5-C5', which decreases the affinity of binding by over 100-fold. Substitution of the C3-G3' base pair showed the next most significant effect, followed by the substitution of the A-T pairs at positions 4 and 6, immediately 3' to the conserved C-G base pairs. A – T substitutions at positions 6-8 also resulted in significant decreases in binding affinity, suggesting that Ndt80

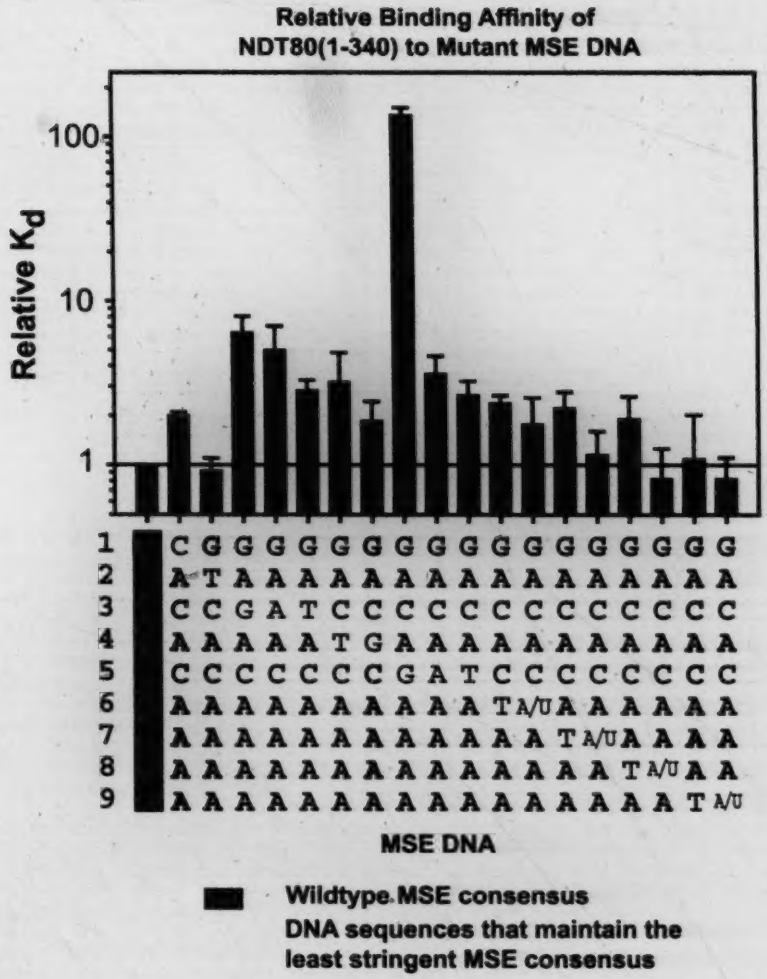
Relative Binding Affinity of
NDT80(1-340) to Mutant MSE DNA
Relative $K_d$
100
10
1
1 C G G G G G G G G G G G G G G G
2 A T A A A A A A A A A A A A A A
3 C C G A T C C C C C C C C C C C
4 A A A A A T G A A A A A A A A A
5 C C C C C C C G A T C C C C C C
6 A A A A A A A A A A T A/U A A A A A
7 A A A A A A A A A A A T A/U A A A A
8 A A A A A A A A A A A A T A/U A A
9 A A A A A A A A A A A A A T A/U
MSE DNA
Wildtype MSE consensus
DNA sequences that maintain the
least stringent MSE consensus

**Figure. 2.8. Mutation of the MSE reduces Ndt80-binding affinity.** The dissociation constants ($K_d$) for the binding of Ndt80(1-340) to wild type and mutated MSE DNAs were determined by electrophoretic mobility shift assay (EMSA). Average $K_d$ values relative to the consensus *SPS4* MSE are plotted, together with standard deviations derived from at least three independent experiments for each mutant. Note the Y - axis is on a logarithmic scale.

specifically recognizes the poly(A)-poly(T) tract at the 3' end of the MSE, and not merely an AT-rich region. To probe further the determinants of specificity within the poly(A)-poly(T) tract, we individually substituted each of the thymine bases within this region with uracils, effectively replacing the 5-methyl group with a hydrogen atom in the modified base. Of these substitutions, only replacement of T6' with U caused a significant decrease in Ndt80-binding affinity, indicating a significant role for the 5-methyl group at this position.

**Overall structure**

We have crystallized and determined the structure of Ndt80(1-340) bound to a 14 base pair DNA duplex containing a consensus MSE sequence derived from the *SPS4* gene, as well as the structure of Ndt80(59-340) in the absence of DNA (Experimental Procedures, Table 2.1, Figure 2.11A). The proteolytically-resistant core domain reveals a central β sandwich structure characteristic of an s-type immunoglobulin (Ig) fold (Bork et al., 1994) (Figure 2.10). An alignment of the free and bound forms of Ndt80 reveals little structural change upon DNA binding (RMSD = 0.8 Å for all Cα atoms in common), other than the ordering of several loops that directly contact DNA (see below). The β-sandwich contains a three-stranded sheet composed of strands a, b and e, packed against a four-stranded sheet composed of strands c', c, f, and g, labeled according to the standard nomenclature (Bork et al., 1994). Each sheet of the β-sandwich contains an additional β-strand, as well as a variety of peripheral secondary structure elements, which are

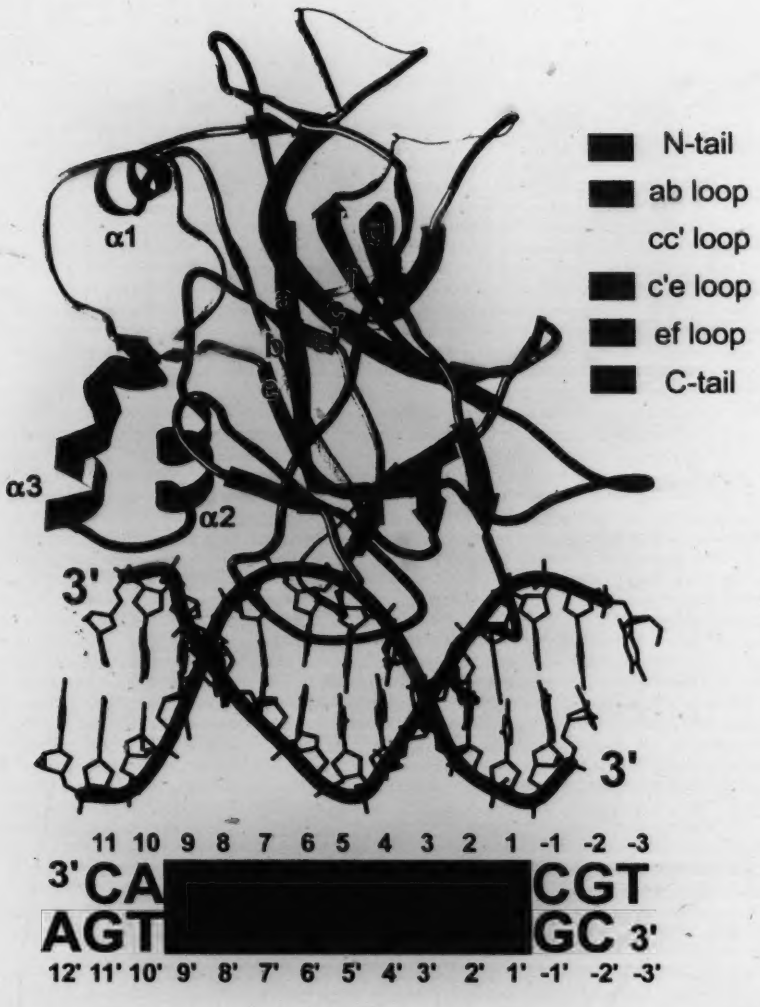


**Figure 2.9. Model of Ndt80(1-340) bound to DNA.** Overview of the Ndt80(1-340) - MSE complex. The regions that contact DNA are rainbow colored in a scheme that is maintained in this chapter. Only the Ig fold strands and α helices are labeled for simplicity.

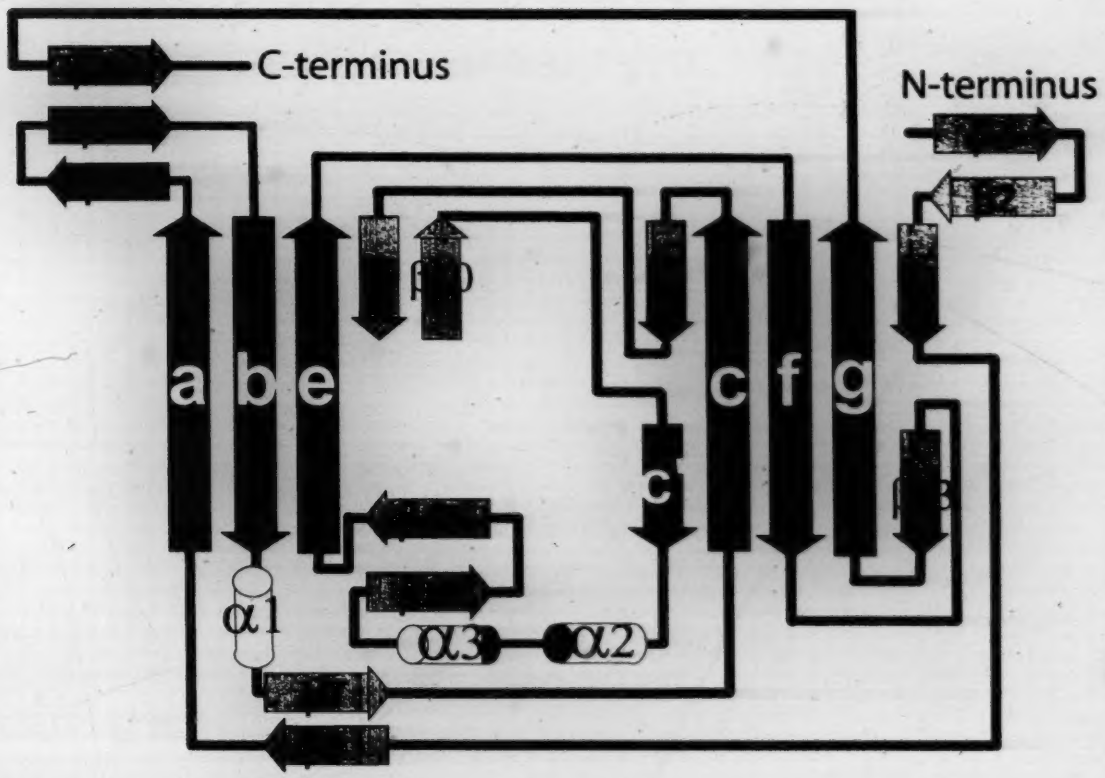


**Figure 2.10 Topology diagram of Ndt80 DNA binding domain.** The conserved s-type Ig fold strands are indicated in black with peripheral strands and helices in gray. The loops involved in DNA binding are color coded as in Figure 2.11.

numbered according to their occurrence in the primary sequence (Figures 2.10, 2.11(B)). A search of the protein structure database using either the DALI (Holm and Sander, 1997) or VAST programs (Madej et al., 1995) revealed that many of the proteins that are most similar to Ndt80 at the structural level are transcription factors that bind DNA through domains containing an s-type Ig fold: members of the p53 (Cho et al., 1994), Rel/NF-κB (Chen et al., 1998a; Chen et al., 1998b; Ghosh et al., 1995; Muller et al., 1995) , STAT (Becker et al., 1998; Chen et al., 1998c), CBF/Runx (Bravo et al., 2001; Tahirov et al., 2001), and T-box families (Muller and Herrmann, 1997) (for a recent review of the Ig-fold family of transcription factors, see (Rudolph and Gergen, 2001). Ndt80 is the first non-metazoan member of the Ig-fold family of transcription factors.

The structure of the complex of Ndt80 bound to the MSE has been refined to 1.4 Å resolution; it reveals a large and complex protein-DNA interface (Figures 2.12, 2.14, and 2.15). Ndt80 interacts with both the major and minor grooves of the DNA through six distinct regions extending from one end of the $\beta$ sandwich. The face of the Ig-fold that contacts DNA is the same as seen in other Ig-fold transcription factors and three of the DNA-contact regions are conserved. The DNA adopts typical B-DNA geometry and has no dramatic bends or kinks. Like the DNA sequences contacted by many of the other Ig-fold transcription factors, the MSE contains a 5' GC-rich region that is specifically recognized through interactions with amino acid side chains via

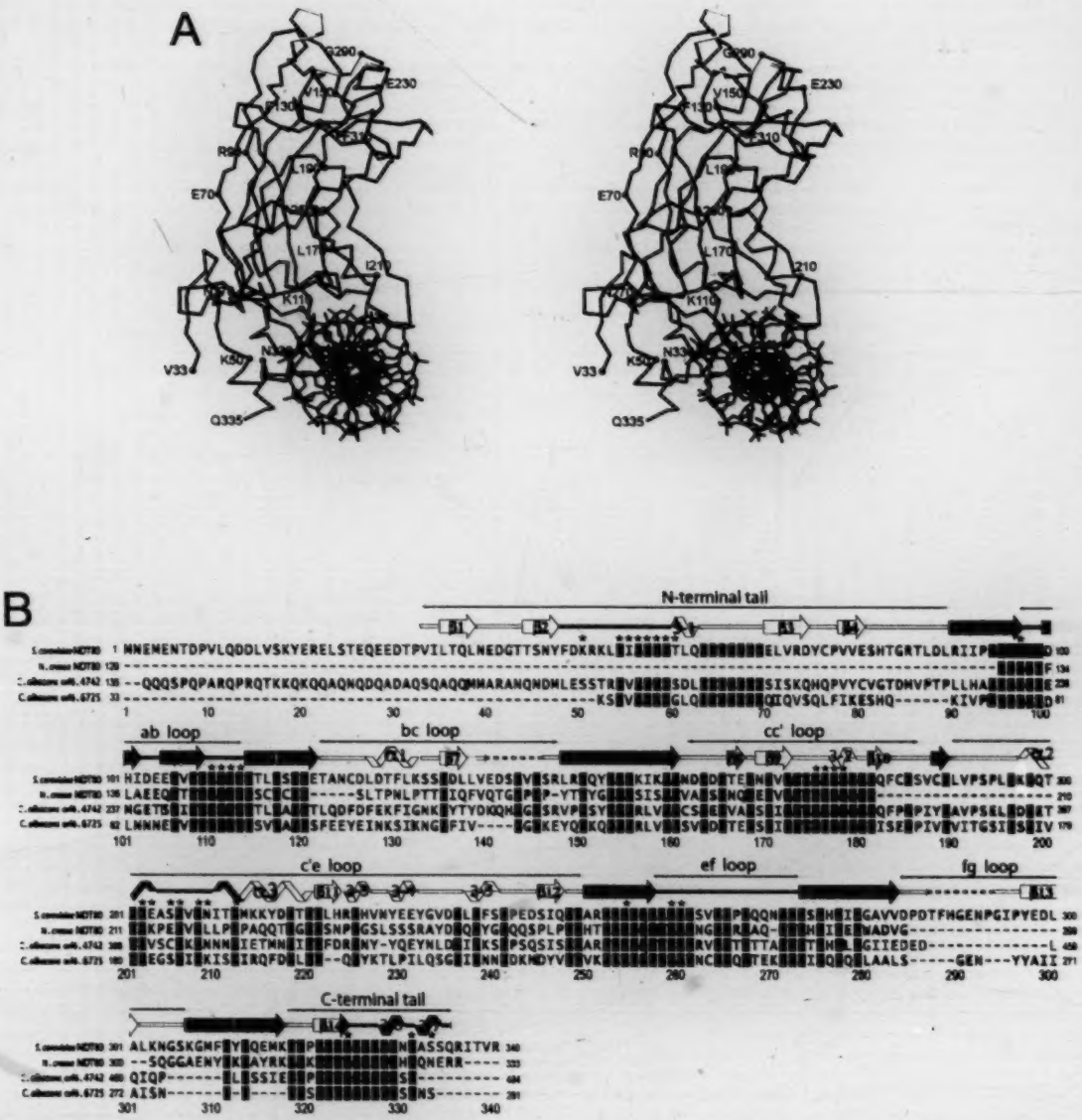


**Figure 2.11. Overview of Ndt80(1-340)-MSE complex. (A)** Cα trace of the complex structure. The green trace represents regions of the protein that can be modeled in both the bound an unbound structures of Ndt80, the red represents regions seen only in the complex (some are disordered in the unbound form others are not physically present), the gray trace represents chain breaks. **(B)** Secondary structure and sequence alignment of the DNA-binding domain of Ndt80. Highlighted regions share ≥75% identity, stars represent residues involved in DNA contacts, red lines represent insertions omitted to maintain an ungapped *S. cerevisiae* sequence.

the DNA major groove, and a 3' AT-rich region that is largely contacted via the DNA minor groove.

**Comparison with other Ig-fold transcription factors**

Not only does Ndt80 share a core β-sandwich topology with the other Ig-fold transcription factors, but it also binds DNA using many of the same loops employed by the other members of the family (Figures 2.9, 2.10, 2.11). For example, the a-b loop in Ndt80 contains a β-hairpin that is oriented perpendicular to the plane of the β-sandwich to interact both with the DNA major groove, and with the N- and C-terminal extensions that also contact DNA. The a-b loop in p53 (Cho et al., 1994), as well as in the NF-κB p50-like proteins (Ghosh et al., 1995; Muller et al., 1995) and NFAT (Chen et al., 1998b) also adopt β-hairpin conformations, although in the case of the NF-κB and NFAT proteins, the hairpin points in the opposite direction to that found in Ndt80, so that more extensive contacts with the DNA can be made. The e-f loop in Ndt80 makes contacts to the backbone of the DNA and, in other Ig-fold transcription factors, this loop plays a similar role, and also interacts directly with the base pairs via the minor groove. The e-f loop is quite variable in size and structure, containing a large helical insert characteristic of the NF-κB p50 family. The C-terminal tail of Ndt80 extends away from the body of the protein to make critical contacts with the DNA major groove. The C-terminal extension of several other Ig-fold transcription factors also interacts with the DNA major groove, although the conformation of this loop is quite variable. For example, in p53, the extension adopts a helical conformation that packs

against the a-b loop and interacts with the DNA major groove, while in NF-κB p50 and AML1/Runx1, the C-terminal tail extends away from the Ig-fold to contact DNA.

Ndt80 also contains several large loop regions that are unique to this protein. Ndt80 contains an N-terminal extension (residues 1-89), which consists of a β-hairpin and a loop that contacts the DNA minor groove and is essential for sequence-specific recognition of the MSE. Residues 1-32 are not visible in the electron density, and these residues are not required for sequence-specific recognition of the MSE (data not shown). The finding that the loop region (residues 50-60) is sensitive to proteolysis in the unbound state but not in complex with DNA suggests that this region is conformationally flexible and adopts its ordered three-dimensional structure only upon binding the MSE. Residues 70-80 constitute a β-strand that is unique to Ndt80 and bonds to the g strand of the c'-c-f-g β-sheet.

The loops linking the b-c, c-c', c'-e and f-g strands of Ndt80 tend to be larger and more complex than equivalent loops found in the other Ig-fold transcription factors. The large c-c' loop contains a β-hairpin structure and makes novel, sequence-specific contacts to the DNA major groove. The c'-e loop contains a helix-loop-helix insert that recognizes the narrow minor groove of the AT-rich portion of the MSE. The f-g loop contains an extra strand (β13) that extends the c'-c-f-g sheet.

Overall, the Ndt80 DNA-binding domain makes more extensive contacts with the MSE than is observed between other, single Ig domains and

their target DNAs. In all, 2600 $Å^2$ of solvent accessible surface area is buried in the Ndt80-DNA interface. This is roughly twice the area buried by other Ig-fold transcription factors. This might explain why many of the other Ig-fold transcription factors only bind DNA with high affinity as dimers (as in the case of the Rel/NF-κB family), or in complex with other, accessory proteins (as is the case for NFAT and AML1/Runx1 proteins) (Rudolph and Gergen, 2001).

**Ndt80-MSE specificity: 5'-YpG-3' recognition**

All Ig-fold transcription factors make sequence-specific contacts to their target DNAs through arginine side chains that recognize the major groove face of guanine bases through a pair of hydrogen bonds between the guanidinium group of the arginine and the N7 and O6 atoms of the guanine (Rudolph and Gergen, 2001). In Ndt80-MSE complexes, the conserved C-G pairs at positions 3 and 5 are recognized in this way, as is the semi-conserved G-C pair at position 1 (Figures 2.12, 2.13). G1 is recognized by R326 from the C-terminal tail, while G3' is recognized by R111 from the a-b loop and G5' is recognized by R177 from the c-c' loop. These contacts explain the fact that mutation of any of the conserved G-C base pairs within the MSE leads to a significant loss of Ndt80-binding affinity (Figure 2.8).

In addition to direct recognition of the G-C base pairs, pyrimidine bases (thymine favoured over cytosine) are conserved at the positions immediately 5' to the guanine nucleotides at positions 3 and 5. Our structure reveals that Ndt80 recognizes the 5' pyrimidine residues by taking advantage of the inherent flexibility of 5'-pyrimidine-purine-3' dinucleotide steps

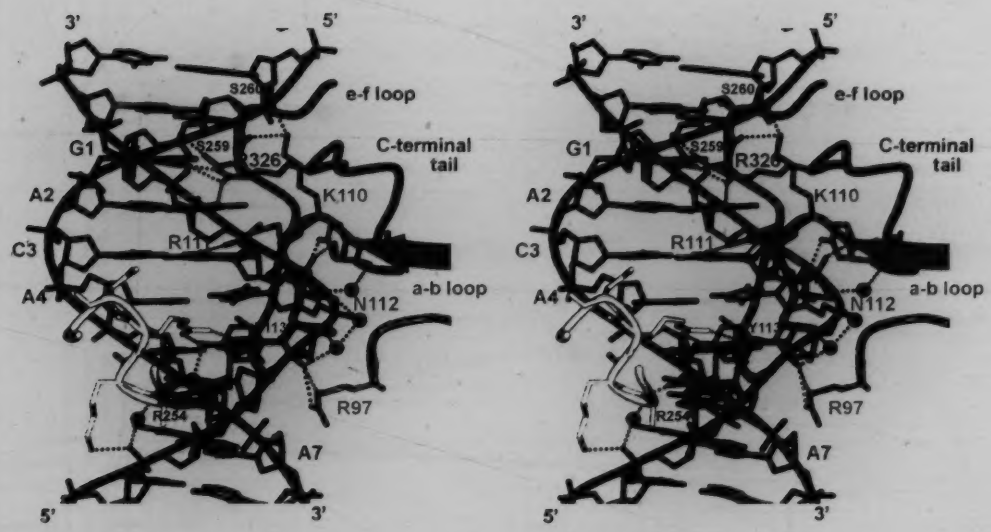


**Figure 2.12 Ndt80-DNA interface in the major groove.** Stereoimage of the major groove interactions between Ndt80 and DNA. The loops involved in this interface include a-b, c-c', and e-f loops as well as the C- terminal tail.

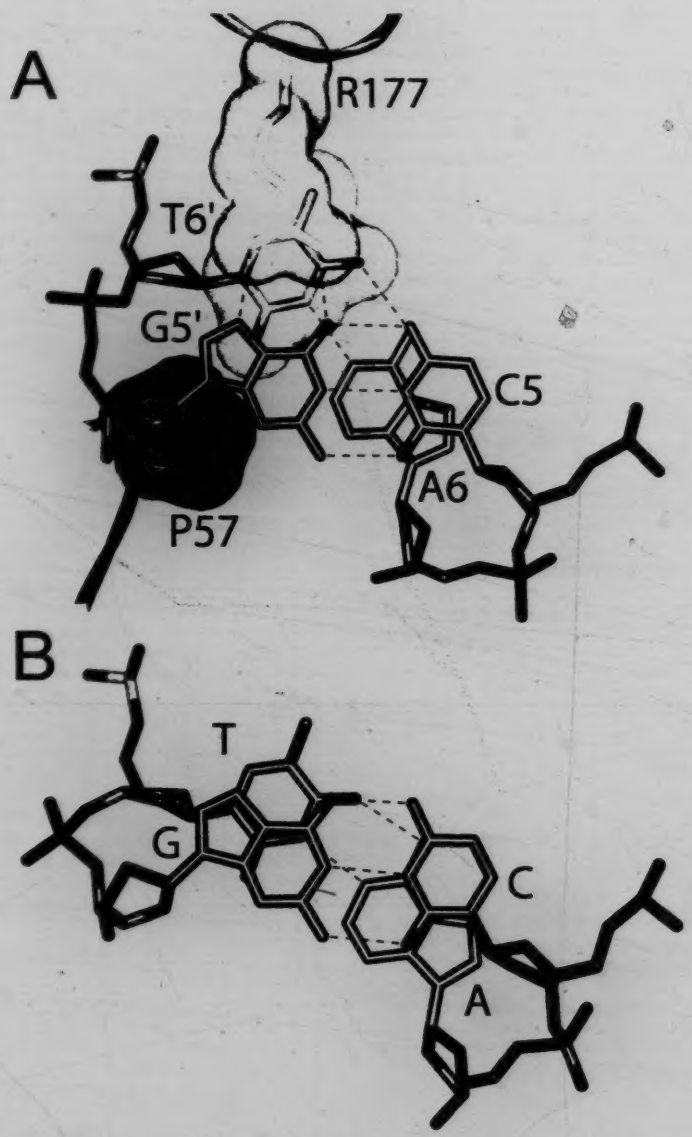


**Figure 2.13 5'-YpG-3' recognition by Ndt80. (A)** View down the DNA helix axis showing base stacking between the CG base pair at position 5 and the AT pair at position 6. Van der Waals surface representations of R177 (yellow), P57 (red) and T6' (gray) are displayed and hydrogen bonding interactions are indicated in green. **(B)** An equivalent CG – AT is shown from the structure of a similar DNA sequence determined in the absence of bound protein (Nelson et al., 1987). The view is such that the CG base pairs in both panels are in identical orientations.

(Olson et al., 1998). This inherent flexibility is probably due to poor overlap between adjacent base pairs in these steps. Comparison of the 5'-T4'pG3'-3' and 5'-T6'pG5'-3' steps from the Ndt80-MSE structure with a 5'-TpG-3' step from the crystal structure of a free DNA of nearly identical sequence (Nelson et al., 1987), reveals that the thymine bases in the Ndt80-bound structure are displaced by ~1.6 Å into the major groove, almost completely sacrificing stacking with the 3' guanine (Figure 2.13). Instead, the thymines hydrogen bonded to the guanine, as well as to the main chain of the arginine and the adjacent residue. This observation explains the fact that mutation of either of the thymines to adenine results in a significant loss of binding affinity to Ndt80 (Figure 2.8). The importance of van der Waals contact between the 5-methyl group of T6' and R177 is demonstrated by the fact that substitution of T6' with U results in a significant loss of Ndt80-binding, whereas T-U substitutions in the other positions of the poly(A)-poly(T) tract do not affect binding. In addition, T6' is also contacted through the minor groove by P57, which also makes significant van der Waals contact with the otherwise exposed face of G5'. The fact that the T6'-G5' step is recognized in a concerted manner via both the major and minor grooves probably explains the dramatic, 100-fold reduction in binding affinity when the C5-G5' base pair is mutated to G-C (Figure 2.8).

Recently, *ab initio* energetic calculations and analysis of protein-DNA structures were used to suggest that arginines might interact most favorably with 5'-purine-guanine-3' dinucleotide steps, where the 3' guanine is

recognized by hydrogen bonding, while the 5' purine interacts with the arginine through cation-$\pi$ interactions (Rooman et al., 2002; Wintjens et al., 2000). In Ndt80, however, a 5' pyrimidine is clearly favored over a 5' purine, probably because the conformational flexibility afforded by the 5'-pyrimidine-guanine-3' allows a greater degree of pyrimidine-arginine contact.

A survey of the structures of other Ig-fold transcription factors bound to their cognate DNAs reveals the AML1 (Runt domain) also employs coupled arginine-guanine hydrogen bonding and pyrimidine stacking to recognize 5'-pyrimidine-guanine-3' steps in its consensus DNA (5'-TGTGGTT-3') (Bravo et al., 2001; Tahirov et al., 2001) . Interestingly, the specific recognition of 5-MeCpG dinucleotide steps over the unmethylated forms by MBD transcriptional repressors is mediated by a pair of arginine residues (Free et al., 2001; Ohki et al., 2001). Whether or not this recognition is also mediated by the coupled hydrogen bonding and stacking interactions between the arginines and the 5-MeCpG dinucleotide awaits the determination of such a structure at high resolution.

**Ndt80-MSE specificity: minor groove recognition**

The poly(A)-poly(T) portion of the MSE (base pairs 6-9) is largely recognized by Ndt80 through interactions with the DNA minor groove (Figures 2.14, 2.15). The N-terminal loop linking $\beta$2 and $\beta$3 enters the minor groove and P57 makes van der Waals contacts with the minor groove face of the A6-

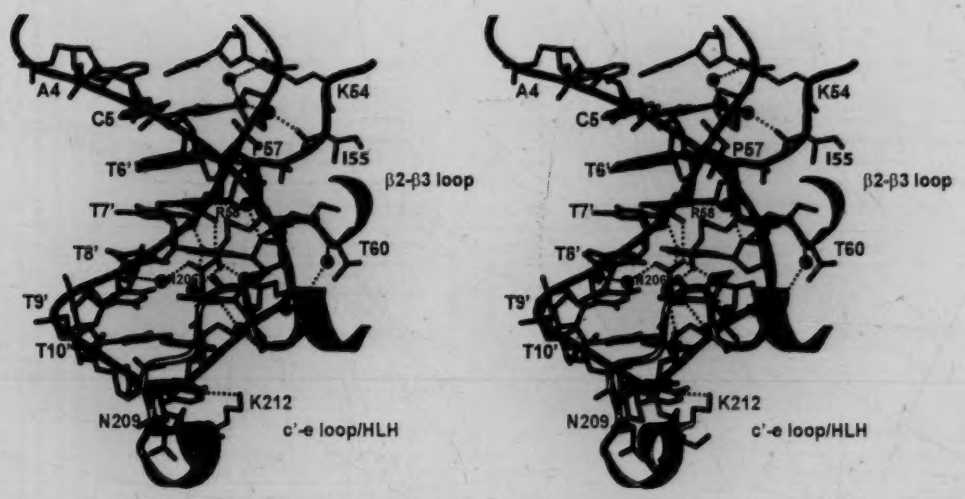


**Figure 2.14 Ndt80-DNA interface in the minor groove.** Stereo image of the minor groove interactions between Ndt80 and DNA. This interface involves the N-terminal tail (β2-β3 loop) and the c'-e loop (HLH).

T6' base pair (Figures 2.14, 2.13(A), 2.15). The next residue in this loop, R58, makes van der Waals contacts to the A7-T7' pair in the minor groove, and the R58 guanidinium group hydrogen bonds with T8', as well as a water that hydrogen bonds to the A9-T9' pair (Figures 2.9, 2.14, 2.15). Substitution of any of these A-T pairs with a G-C pair would result in a steric clash between the 2-amino group of the guanine base and the side chain. As a result of these interactions, the DNA minor groove near this end of the poly(A)-poly(T) region is significantly wider than that normally seen in B-DNA (~13 Å inter-strand phosphate distance, compared with an ~11 Å distance in B-DNA). However, toward the 3' end of the poly(A)-poly(T) tract, the minor groove narrows significantly to ~9.5 Å, similar to the minor groove width observed in the crystal structures of poly(A)-poly(T) in the absence of protein (Nelson et al., 1987). The narrow minor groove in the vicinity of base pairs A7-T7' – C11-G12' is recognized by residues extending from a helix-loop-helix motif present in the c'-e loop (Figure 2.14). These residues make a number of direct and solvent-mediated hydrogen bond and electrostatic contacts to both strands across the narrow minor groove.

While our structure explains why A-T pairs are favored over G-C pairs in the 3' region of the MSE, it is less clear why substitution of any A-T pair with a T-A at positions 7-9 is unfavorable (Figure 2.8). Such substitutions would have little affect on either the disposition of hydrogen bonding groups or the width of the minor groove. A possible explanation may lie in the fact that poly(A)-poly(T) tracts are more rigid than A-T DNAs containing 5'-TpA-3'

dinucleotide steps (Suzuki et al., 1996). Because of this, an additional entropic cost might be associated with Ndt80 binding to a flexible, mixed A-T sequence, over a more rigid poly(A)-poly(T) tract that is pre-set in a conformation appropriate for Ndt80 binding. The crystal structure of a free DNA with a sequence that is almost identical to the MSE provides an approximation of the conformation of the MSE DNA in its free state (Nelson et al., 1987). While the structure of this DNA is very similar to the conformation of the MSE DNA bound to Ndt80, interesting differences are present. In addition to the differences in minor groove width mentioned above, the poly(A)-poly(T) tract in its unbound form also exhibits pronounced propeller twisting that at once enhances intra-strand base stacking and hydrogen bonding between adenine and thymine bases in adjacent base pairs. This propeller twisting is not observed in the Ndt80-bound form of the MSE, and suggests that protein binding may alter the conformation of the DNA in subtle but significant ways that ultimately modulates the binding affinity.

**Ndt80-MSE specificity: backbone contacts**

In addition to the contacts described above that play a key role in the sequence-specific recognition of the MSE, Ndt80 displays a complex network of electrostatic and hydrogen bonding interactions that anchor the protein to the DNA (Figures 2.12, 2.14, 2.15). The N-terminal and C-terminal extensions, the a-b, c-c', c'-e, and e-f loops, which mediate sequence-specific interactions with the DNA, all make contacts to the DNA backbone,

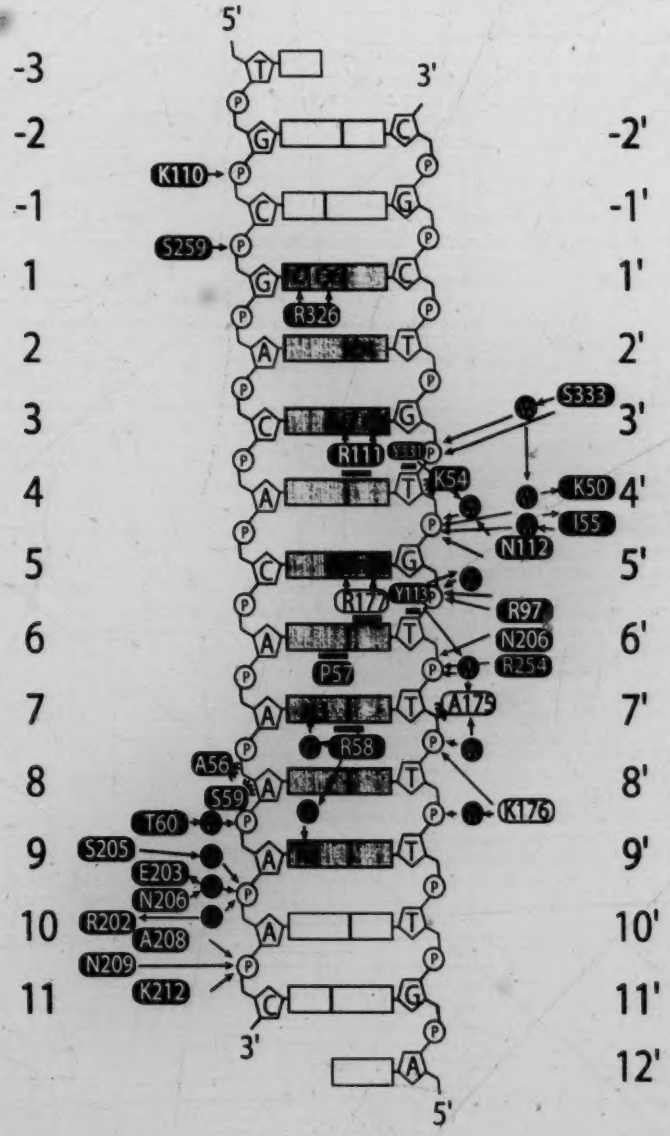


**Figure 2.15 Schematic of Ndt80-DNA interactions.** The MSE bases are highlighted in blue and numbered from 1-9 with the complementary strand distinguished with a prime. Protein residues are colored according to the scheme laid out in Figure 2.9. Electrostatic and polar interactions are represented with arrows, while van der Waals contacts are indicated with hash marks. The blue circles represent water molecules that are highly ordered within the interface.

many of which are mediated by well-ordered solvent molecules.

**Ndt80 flexibility and DNA binding**

A comparison of the DNA bound form of Ndt80(1-340) with the unbound form of Ndt80(59-340) indicates that several of the DNA contact regions only become structured upon interactions with DNA (Figure 2.11(A)). We observed no interpretable electron density for either the c-c' loop (corresponding to residues 175-184), or the C-terminal tail (residues 326-334) in the unbound form of Ndt80, indicating that these critical DNA-contact regions are flexible in the absence of DNA. However, the a-b, c-c', helix-loop-helix, and e-f loops are ordered and adopt conformations that are essentially identical to those found in the DNA-bound form. In particular, R111 from the a-b loop that recognizes the 5'-TpG-3' dinucleotide step at position 3'-4' adopts a side chain geometry almost identical to that seen in the DNA-bound form. Thus, these loops appear to be pre-aligned for interactions with the DNA, while the N- and C-terminal extensions, as well as the c-c' loop, are more flexible and probably only adopt stable structures when complexed with DNA.

**Interactions between Ndt80 and other transcription factors**

While Ndt80 is the key transcription factor that ultimately allows the continuation of meiosis after the successful completion of recombination,

other factors must also influence middle gene expression. Evidence for this comes from the finding that many middle genes are transcriptionally activated in Ndt80-deficient cells during meiosis, albeit to a lower level than in isogenic strains containing wild-type Ndt80 (Chu et al., 1998). Part of this Ndt80-independent induction may be due to the release of repression imposed on middle genes by Sum1-Hst1 or other repressor complexes during vegetative growth and the early stages of meiosis (Pijnappel et al., 2001; Xie et al., 1999). Sum1 has been shown to interact with MSEs in a sequence-specific manner, possibly utilizing its two proposed AT-hook peptide motifs (Aravind and Landsman, 1998), which likely make sequence-specific contacts with the A/T-rich portion of the MSE. Because Ndt80 makes extensive contacts with both the backbone and floor of the minor groove of the A/T-rich region, it is very likely that the Ndt80 activator and the Sum1 repressor compete for the same DNA regulatory elements. This competition may ultimately help to regulate the precise timing and level of expression of key middle genes (Lindgren et al., 2000; Xie et al., 1999).

Detailed examination of the DNA regulatory elements that activate middle genes have indicated that, while the MSEs of these genes are essential to their correct developmental regulation, other neighboring sequences play an important role in enhancing MSE-dependent activation. For example, binding sites for the general transcriptional regulator ABF1 are occasionally found adjacent to MSEs and these sites have been shown in some cases to enhance MSE-dependent activation ~10-fold (Hepworth et al.,

1995; Ozsarac et al., 1997). The close proximity of these sites to the MSEs suggests that Ndt80 and ABF1 might act cooperatively to activate transcription.

## REFERENCES


Aravind, L., and Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res *26*, 4413-4421.

Becker, S., Groner, B., and Muller, C. W. (1998). Three-dimensional structure of the Stat3beta homodimer bound to DNA. Nature *394*, 145-151.

Bork, P., Holm, L., and Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol *242*, 309-320.

Bravo, J., Li, Z., Speck, N. A., and Warren, A. J. (2001). The leukemia-associated AML1 (Runx1)--CBF beta complex functions as a DNA-induced molecular clamp. Nat Struct Biol *8*, 371-378.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., *et al.* (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr *54 ( Pt 5)*, 905-921.

Chen, F. E., Huang, D. B., Chen, Y. Q., and Ghosh, G. (1998a). Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA. Nature *391*, 410-413.

Chen, L., Glover, J. N. M., Hogan, P. G., Rao, A., and Harrison, S. C. (1998b). Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. Nature *392*, 42-48.

Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell, J. E., Jr., and Kuriyan, J. (1998c). Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. Cell *93*, 827-839.

Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346-355.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Chu, S., and Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. Mol Cell *1*, 685-696.

Collaborative Computational Project, N. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr *50*, 760-763.

Cowtan, K. (1994), In Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, pp. 34-38.

Free, A., Wakefield, R. I., Smith, B. O., Dryden, D. T., Barlow, P. N., and Bird, A. P. (2001). DNA recognition by the methyl-CpG binding domain of MeCP2. J Biol Chem *276*, 3353-3360.

Ghosh, G., van Duyne, G., Ghosh, S., and Sigler, P. B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. Nature *373*, 303-310.

Hepworth, S. R., Ebisuzaki, L. K., and Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in Saccharomyces cerevisiae. Mol Cell Biol *15*, 3934-3944.

Hepworth, S. R., Friesen, H., and Segall, J. (1998). NDT80 and the meiotic recombination checkpoint regulate expression of middle sporulation-specific genes in Saccharomyces cerevisiae. Mol Cell Biol *18*, 5750-5761.

Holm, L., and Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res *25*, 231-234.

Jolly, E., Chin, C. S., Herskowitz, I., and Li, H. (2005). Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis. BMC Bioinformatics *6*, 275.

Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A *47 ( Pt 2)*, 110-119.

Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C., and Glover, J. N. M. (2002). Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. Embo J *21*, 5721-5732.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst, 283-291.

Lindgren, A., Bungard, D., Pierce, M., Xie, J., Vershon, A., and Winter, E. (2000). The pachytene checkpoint in Saccharomyces cerevisiae requires the Sum1 transcriptional repressor. Embo J *19*, 6489-6497.

Madej, T., Gibrat, J. F., and Bryant, S. H. (1995). Threading a database of protein cores. Proteins *23*, 356-369.

Muller, C. W., and Herrmann, B. G. (1997). Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. Nature *389*, 884-888.

Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., and Harrison, S. C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. Nature *373*, 311-317.

Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. Cell *105*, 487-497.

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., and Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A *95*, 11163-11168.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode, In Methods in Enzymology, C. W. Carter Jr., and R. M. Sweet, eds. (New York: Academic Press), pp. 307-326.

Ozsarac, N., Straffon, M. J., Dalton, H. E., and Dawes, I. W. (1997). Regulation of gene expression during meiosis in Saccharomyces cerevisiae: SPR3 is controlled by both ABFI and a new sporulation control element. Mol Cell Biol *17*, 1152-1159.

Perrakis, A., Morris, R., and Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. Nat Struct Biol *6*, 458-463.

Pijnappel, W. W., Schaft, D., Roguev, A., Shevchenko, A., Tekotte, H., Wilm, M., Rigaut, G., Seraphin, B., Aasland, R., and Stewart, A. F. (2001). The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. Genes Dev *15*, 2991-3004.

Rooman, M., Lievin, J., Buisine, E., and Wintjens, R. (2002). Cation-pi/H-bond stair motifs at protein-DNA interfaces. J Mol Biol *319*, 67-76.

Rudolph, M. J., and Gergen, J. P. (2001). DNA-binding by Ig-fold proteins. Nat Struct Biol *8*, 384-386.

Suzuki, M., Yagi, N., and Finch, J. T. (1996). Role of base-backbone and base-base interactions in alternating DNA conformations. FEBS Lett *379*, 148-152.

Tahirov, T. H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., *et al.* (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell *104*, 755-767.

Terwilliger, T. C. (2000). Maximum-likelihood density modification. Acta Crystallogr D Biol Crystallogr *56 ( Pt 8)*, 965-972.

Terwilliger, T. C., and Berendzen, J. (1999). Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr *55 ( Pt 4)*, 849-861.

Wintjens, R., Lievin, J., Rooman, M., and Buisine, E. (2000). Contribution of cation-pi interactions to the stability of protein-DNA complexes. J Mol Biol *302*, 395-410.

Xie, J., Pierce, M., Gailus-Durner, V., Wagner, M., Winter, E., and Vershon, A. K. (1999). Sum1 and Hst1 repress middle sporulation-specific gene expression during mitosis in Saccharomyces cerevisiae. Embo J *18*, 6448-6454.

# Chapter 3:

## A new mode of 5'-YpG-3' step recognition by amino acids:

## Coupled hydrogen bond and stacking interactions

## SUMMARY

The combined biochemical and structural studies of hundreds of protein-DNA complexes has indicated that sequence-specific interactions are mediated by two mechanisms termed direct and indirect readout. Direct readout involves direct interactions between the protein and base-specific atoms exposed in the major and minor grooves of DNA. For indirect readout, the protein recognizes DNA by sensing conformational variations in the structure dependent on nucleotide sequence, typically through interactions with the phosphodiester backbone. Based on the structure of NDT80 bound to DNA (described in the previous chapter) in conjunction with a search of the existing PDB database, we propose a new adaptation of sequence-specific recognition that uses both direct and indirect readout. In this mode, a single amino acid side chain recognizes two consecutive base pairs. The 3'-base is recognized by canonical direct readout, while the 5'-base is recognized through a variation of indirect readout, whereby the conformational flexibility of the particular dinucleotide step, namely a 5'-pyrimidine-purine-3' step, facilitates its recognition by the amino acid via cation - $\pi$ interactions. In most cases, this mode of DNA recognition helps to explain the sequence specificity of the protein for its target DNA. The majority of this work was published in JMB (Lamoureux et al., 2004).

## INTRODUCTION

The structure of Ndt80 bound to an MSE - containing DNA described in the previous chapter was determined at 1.4 Å resolution (Lamoureux et al., 2002), thereby providing an excellent opportunity to understand in detail the way in which this protein selectively binds the MSE (5'-gNCRCAAAW-3') (Hepworth et al., 1995; Ozsarac et al., 1997). The poly(A)-poly(T) region is recognized through the minor groove while the CG rich 5' end is recognized by three arginine residues that make bidentate hydrogen bonds to the major groove face of the C-G basepairs. While these interactions are quite similar to the ways in which other transcription factors recognize DNA, the structure revealed an unexpected mode of recognition of the conserved pyrimidine residues immediately 5' to the guanine residues at positions 3 and 5.

Ndt80 recognizes the intrinsic flexibility of these 5'-YpG-3' steps through the hydrogen bonding of an arginine sidechain to the major groove face of the guanine base coupled to the stacking of the 3' pyrimidine ring on the co-planar guanidinium group of that same arginine. In general, 5'-YpR-3' steps are significantly more flexible than other dinucleotide steps, probably due to the low degree of base pair overlap within the 5'-YpR-3' step (Olson et al., 1998). As a result, it is less energetically costly to deform these dinucleotide steps, and 5'-YpR-3' steps are often sites of DNA bending (Dickerson and Chiu, 1997; Olson et al., 1998). Interestingly, the 5'-YpR-3' steps presented here are not sites of significant bending. The deformation of the 5'-YpR-3' steps in Ndt80 is accompanied by a shift in the backbone

conformation of the 5'-pyrimidine and/or its complementary purine from the common BI conformation, which is characterized by ε and ζ torsion angles in the (t/g-) range, where ε-ζ ≈ -90°, to the less common BII conformation, where ε, ζ = (g-/t) and ε-ζ ≈ +90°. BII conformations have long been known to lead to base unstacking (Prive et al., 1987) and theoretical studies predict that these transitions will be most prevalent in the flexible 5'-YpG-3' dinucleotide steps (Bertrand et al., 1998).

The importance of the 5'-pyrimidine within both of these 5'-YpG-3' sequences for Ndt80 binding was demonstrated by the finding that mutation of either 5'-pyrimidine to a purine resulted in a 3-5 fold reduction in DNA binding affinity (Fingerman et al., 2004; Lamoureux et al., 2002). Moreover, mutation of the conserved thymine at position 6 to a uracil, also caused a significant (2-fold) reduction in binding affinity, consistent with the idea that the 5-methyl group of thymine is critical for stacking interactions with the arginine (Lamoureux et al., 2002). To ascertain if this mode of recognition is utilized by other proteins, we searched the protein-DNA database with a PERL script that crudely parsed the pdb's files looking for similar interactions. Amazingly we found examples of this type of hydrogen bond coupled to cation - π interactions in nearly all of the major classes of transcription factor. In addition, in many cases where these interactions are seen they also explain hitherto unexplained DNA sequence preferences.

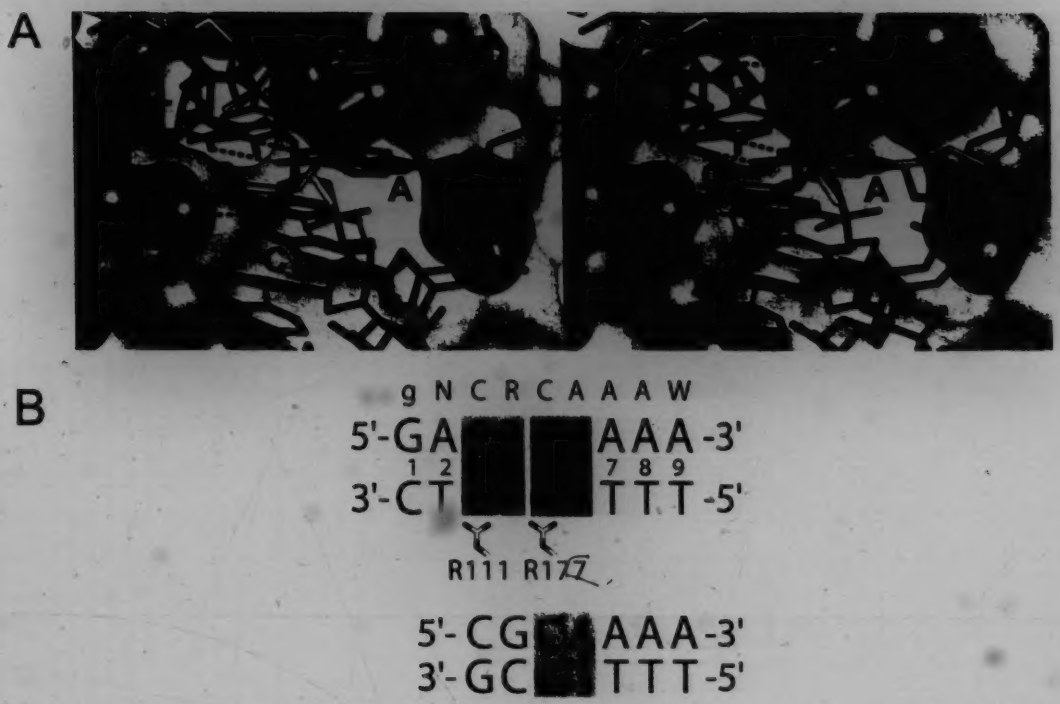


**Figure 3.1. Alignment of Ndt80-DNA complex with reference DNA structure**. **(A)** The NDT80-DNA complex (blue) is aligned to the reference DNA structure (orange) using the 3' G-C base pair of the 5'-YpG-3' step. The surface of Ndt80 is shown in transparent grey with a green stick representation for Arg177 and Pro57 involved in the recognition of the 5'-YpG-3' step. Note that the base pairs 5' to the shifted thymine align well with the reference DNA indicating that this distortion is limited to the 5'-YpG-3' step. A similar distortion of the 5'-YpG-3' step is seen at the 5'-TpG-3' step contacted by Arg111. **(B)** Consensus MSE sequence aligned to the actual sequence used in the crystal structure. The two unstacked 5'-TpG-3' steps are on the bottom strand and are highlighted. The bottom duplex is the reference DNA structure (PDB ID: **1D98**) with the 5'-TpG-3' step highlighted.

## EXPERIMENTAL PROCEDURES

### Parsing the database

The 553 protein-DNA structures were obtained from the September 15, 2003 release of the protein database (http://www.rcsb.org/pdb/). The database contains all structures solved by x-ray crystallography with a resolution of 3.0 Å or better containing both DNA and protein. See Appendix A for a listing of the PDB files used in this chapter based on the above criteria. A Perl script (ArgStack.pl, see Appendix B) was used to find all 5'-YpG-3' steps with an arginine sidechain within hydrogen bonding distance of the O6 and/or N7 of the guanine and simultaneously in close proximity to the 5'-pyrimidine (as determined by arginine CZ to pyrimidine C5 distance in the case of arginine residues). The distance cutoffs used for the hydrogen bonding distance and proximity to 5'-pyrimidine were 3.0 and 6.0 Å, respectively.

### Assessing unstacking

Next, the degree of stacking of the 5'-pyrimidine on the 3'-purine was assessed using the criteria described in Figure 3.2. We considered a 5'-YpG-3' step unstacked if both the $^{Y}$C2-$^{G}$C5 and $^{Y}$O2-$^{G}$C4 interatomic distances were longer than the $^{Y}$O2-$^{G}$C5 distance. Typical values for these distances are 3.7 Å for $^{Y}$C2-$^{G}$C5, 3.8 Å for $^{Y}$O2-$^{G}$C4 , and 3.9 Å for $^{Y}$O2-$^{G}$C5. 5'-YpG-3' steps with any of these distances greater than 5 Å were removed to eliminate non-B DNA structures from the analysis. In this way, 269 5'-YpG-3' steps in contact with arginine residues were retrieved from the database, of which 81

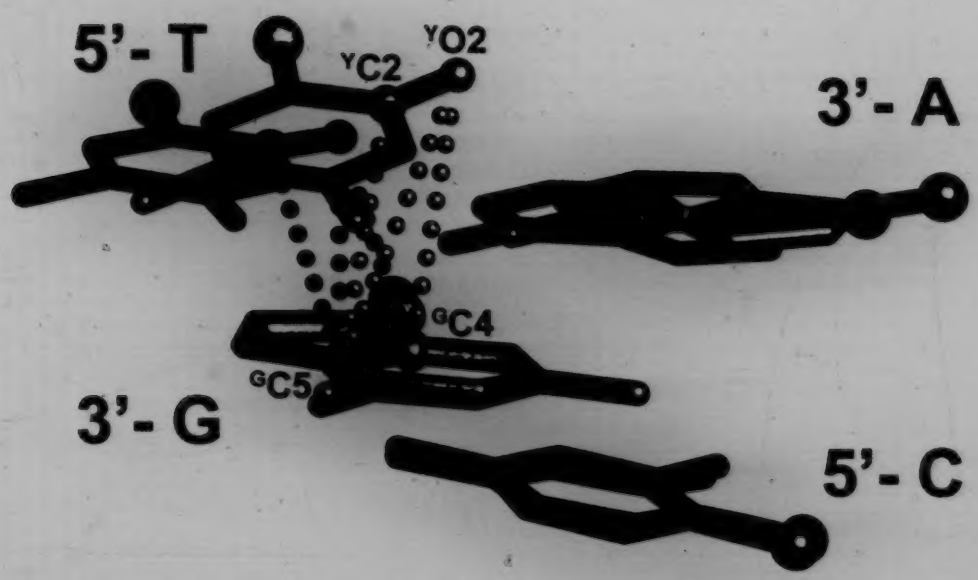


**Figure 3.2 Criteria for identifying unstacked 5'-YpG-3' steps.** This diagram shows the YpG bases as stick representations and the C1 atom as a sphere. Typical B-form YpG DNA is shown in orange and an unstacked YpG from the Ndt80 complex is shown in blue. Three interatomic distances were measured; $^{Y}$O2-$^{G}$C4, $^{Y}$C2-$^{G}$C5, and $^{Y}$O2-$^{G}$C5, where the superscript text indicates the identity of the base. In standard B-DNA, the $^{Y}$O2-$^{G}$C4 and $^{Y}$C2-$^{G}$C5 distances are approximately equal while the $^{Y}$O2-$^{G}$C5 distance is greater. In the Ndt80-DNA complex, the pyrimidine shifts into the major groove and the $^{Y}$O2-$^{G}$C4 and $^{Y}$C2-$^{G}$C5 distances become larger while $^{Y}$O2-$^{G}$C5 is shortened. We consider a dinucleotide step to be unstacked if the $^{Y}$O2-$^{G}$C5 distance is shorter than both the $^{Y}$O2-$^{G}$C4 and the $^{Y}$C2-$^{G}$C5 distances.

(or 30.1%) are unstacked. We next repeated the search using a non-redundant database in which structures with a sequence homology of 90% or greater were removed. The non-redundant database, containing 209 structures, contained 142 5'-YpG-3' steps in contact with arginine, of which 30 (or 21%) are unstacked. In a similar fashion we assessed the stacking for all 5'-YpG-3' steps, regardless of proximity to amino acid side chains, in both the redundant and non-redundant databases. 428/2124 (or 20%) of 5'-YpG-3' steps in the redundant dataset and 176/907 (or 19%) of 5'-YpG-3' steps in the non-redundant dataset were unstacked by our criteria. Finally, a DNA-only database was extracted from the RCSB database, which consisted of all structures containing only DNA, solved by x-ray crystallography to 3.0 Å resolution or better, and identified as B-DNA in the PDB header. 142/1444 (or 10%) of 5'-YpG-3' steps of the DNA-only database are unstacked.

**Aligning to a reference DNA**

The structures retrieved from this automated procedure were then inspected visually to ascertain if the steps are unstacked independently of the criteria above. Each structure was aligned to the reference DNA (PDB ID: **1D98**) by least squares superposition (as implemented in O (Jones et al., 1991)) of the 3'-purine base and its Watson-Crick partner onto the guanine-cytosine pair of the reference step (residue B22 (G) and A3 (C) of the reference). The displacements of the centroids of the 5'-pyrimidine base and its base-paired partner relative to the reference structure were then

determined. For the protein-DNA complexes for which there are multiple independent structures deposited in the RCSB database, we have only discussed those structures for which unstacking is consistently observed in a majority of the deposited structures.

**DNA helical parameter calculations**

DNA helical parameters and torsion angles of the extracted structures were calculated using 3DNA (Lu et al., 2000). The energy of displacement of the helical parameters from their mean values, expressed in terms of $k_BT/2$, was calculated as described (Olson et al., 1998). The square root of this value gives the number of standard deviations from the minimum energy represented by the average parameters (i.e. Z-score).

Similar searches were also performed for 5'-YpR-3' steps in contact with histidine, glutamine, lysine or asparagine via the major groove. Significant unstacking was only observed for 5'-YpG-3' steps in contact with histidine, but not for any of the other side chains.

## RESULTS AND DISCUSSION:

### Database search for interactions between arginine and 5'-YpG-3'

We searched a database of structures of proteins-DNA complexes for 5'-YpG-3' dinucleotide steps in which an arginine side chain is simultaneously hydrogen-bonded to the guanine and in van der Waals contact with the 5'-pyrimidine which is shifted into the major groove. The criteria used to determine whether the pyrimidine was shifted into the major groove is summarized in Figure 3.2. The search of 553 protein-DNA complexes derived from the protein database (http://www.rcsb.org/pdb/) uncovered 13 distinct complexes which clearly display this kind of interaction (Table 3.1, 3.2 see Experimental Procedures). One of these proteins, the AML1/Runt domain, is structurally related to Ndt80. AML1/Runt, like Ndt80, is an Ig-fold transcription factor, and binds DNA using the same edge of the Ig-fold β-sandwich (Bravo et al., 2001; Tahirov et al., 2001). The other proteins uncovered in the search have a variety of unrelated structures. MATα2, Pbx, and Ubx are all homeobox proteins, while the *E. coli* PurR repressor and CRE recombinase recognize DNA via helix-turn-helix motifs. E2F and DP proteins utilize a winged-helix motif to bind their target sequence, while ZIF268 recognizes DNA using 3 tandem Zn fingers each of which recognize 3 consecutive base pairs. C/EBPβ is a homodimeric bZIP transcription factor that contacts DNA using two highly positively charged α-helices. The *E. coli* replication terminator protein Tus recognizes DNA using inter-domain β-strands that contact a deformed DNA major groove.

**Conformational analysis of unstacked 5'-YpG-3' steps**

To compare the conformations of the DNA in these complexes, we aligned each structure on a reference, unbound DNA structure (PDB ID:1D98) (Nelson et al., 1987), the sequence of which is nearly identical to the MSE. The 5'-YpG-3' step from this DNA that we have used for the reference structure is very similar to an averaged 5'-YpR-3' step as derived from the DNA structure database (Olson et al., 1998). With the structures aligned in this way, the 5'-pyrimidine bases are all displaced into the major groove by distances between 0.7 and 2.9 Å. To maintain pairing with the shifted pyrimidine, the complementary purine slides between 0.5 and 2.9 Å along its long axis (Figure 3.3). In general, the inherent stacking symmetry of the 5'-YpG-3' step is broken such that base-base stacking is reduced in the strand that is contacted by the arginine, while it is maintained in the complimentary strand. While these base displacements are in general quite large, all torsion angles remain in their most favored positions for B-DNA. None of these structures adopt a true BII conformation ($\varepsilon$-$\zeta$ > 50°) like that seen in Ndt80. Nevertheless, a large proportion have $\varepsilon$-$\zeta$ values between 0 and 40° for the 5'-pyrimidine. This is a large deviation from the mean BI value of -80° ± 40 and may indicate a shift toward a BII conformation. Because of the difficulty in accurately modeling the phosphate backbone at the moderate resolutions of most of these structures, it is possible that some may indeed adopt a BII conformation. In general, the positions of the phosphates of both the guanine

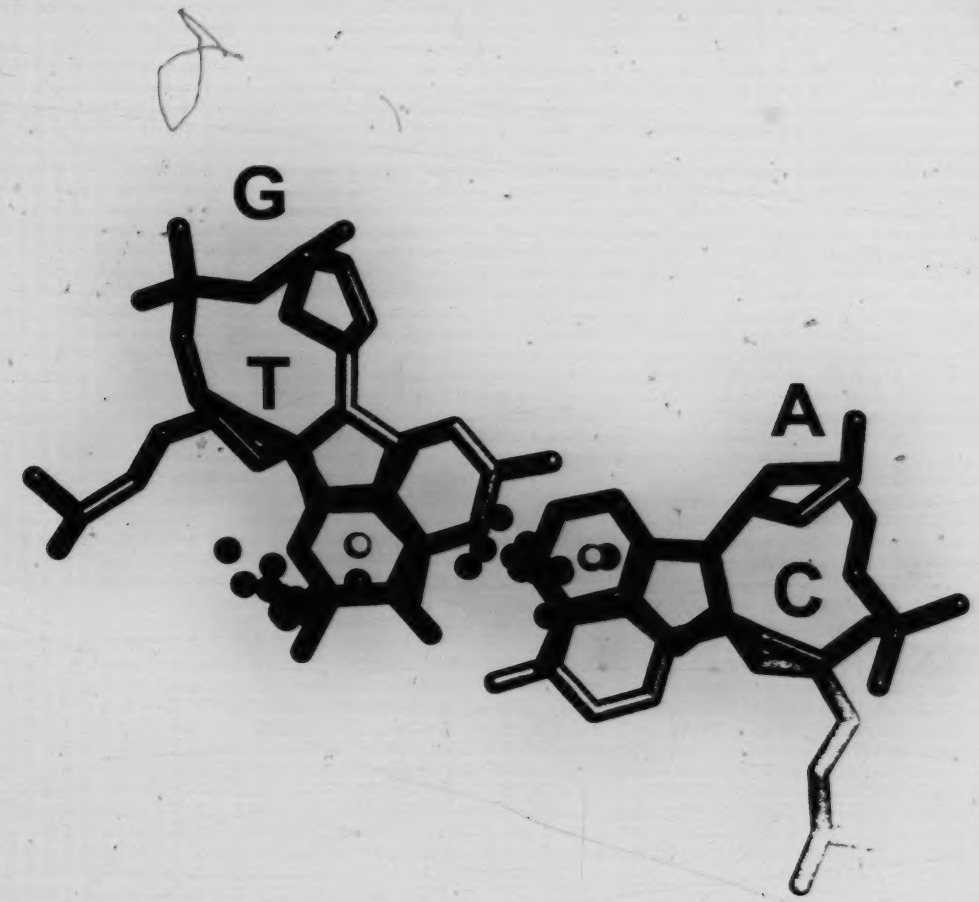


**Figure 3.3 Summary of base displacements.** The 5'-pyrimidine (red) and its complementary purine (green) of the reference DNA are shown viewed down the helical axis. The centroid of the pyrimidine ring and the centroid of the 6 member ring of the purine are shown as yellow spheres for the reference DNA. The corresponding centroids for the average 5'-YpG-3' step, as calculated by 3DNA, are displayed as blue spheres, while the centroids of the shifted pyrimidine rings are shown as red spheres with the complementary purine centroids as green spheres.

and the 5'-pyrimidine are also shifted towards the major groove. In each of the complexes, the shifted phosphates are contacted by the protein through either salt bridges or hydrogen bonding interactions. This suggests that the 5'-YpG-3' deformation not only enhances stacking with the arginine, but also facilitates backbone interactions that may be critical for specificity through indirect readout. Conversely it is also possible that these backbone interactions may facilitate the 5'-YpG-3' deformation.

We have also analyzed the DNA conformation in terms of the 6 independent parameters that fully describe the conformation of two successive base pairs within a double stranded DNA structure (Dickerson et al., 1989) (Table 3.1, 3.2). Intriguingly, most of the 5'-YpG-3' steps show significant negative shift and negative tilt, but no consistent trend away from standard values is observed for any of the other parameters. The negative shift corresponds to a movement of the 5'-pyrimidine into the major groove. The negative tilt corresponds to a change in the angle between the adjacent base pairs of the 5'-YpG-3' step such that the bases in contact with the arginine on one strand have a smaller rise than their complimentary bases on the opposite strand. This negative tilt is allowed because of the low degree of stacking between bases of the 5'-YpG-3' step in contact with the arginine.

We also estimated the energy cost associated with each of these base pair steps as derived from their helical parameters. The costs vary from 2.3 to

**Table 3.1 Summary of database search and DNA parameters for YpG steps contacted by arginines**

| PDB ID | Fold Type | DNA complex | Resol-ution | Arg residue [1] | DNA consensus | Y to ref Y distance [2] | Shift | Slide | Rise | Tilt | Roll* | Twist | Energy [3](KT/2) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1AKH | Homeo-domain | MAT **a1alpha2** | 2.50 | R54 (B185) | **TG**T | 2.07 | -0.06 | -0.79 | 3.51 | -1.23 | 9.82 | 41.1 | 7.1 |
| 1AKH | | MAT **a1**alpha2 | 2.50 | R55 (A124) | **TG**ATGT | 0.70 | 0.32 | 0.11 | 3.02 | -0.7 | 5.7 | 35.9 | 2.4 |
| 1B72 | | **Pbx1**-Hox1 | 2.35 | R55 (B290) | R**TG**ATT | 2.02 | -0.83 | 0.42 | 3.34 | -3.75 | 9.01 | 33.9 | 4.5 |
| 1B8I | | Ubx-**Exd** | 2.40 | R55 (B258) | R**TG**ATT | 1.58 | -0.66 | 0.36 | 3.45 | -1.96 | 10.8 | 35.1 | 3.6 |
| 1QPZ | Helix-turn-helix motif | PurR | 2.50 | R26 (A26) | A**YG**CAAAC | 2.25 | -1.49 | -0.67 | 3.56 | 1.15 | -6.52 | 34.4 | 17.1 |
| 5CRX | | Cre | 2.70 | R259 (A259) | TATA**CG**AAGTTAT | 1.78 | -1.22 | 0.3 | 3.04 | -1.51 | 0.99 | 32.8 | 5.7 |
| 1H9D | Ig-fold Beta sandwich | **AML1**-CBF$\beta$ | 2.60 | R174 (C174) | YG**YG**GTY | 2.11 | -1.24 | -0.18 | 3.25 | -4.45 | 2.28 | 33.7 | 4.9 |
| 1MNN | | NDT80 | 1.40 | R111 (A111) | WTTTG**YG**Nc | 2.37 | -0.95 | -0.3 | 3.73 | -5.47 | 3.98 | 39.8 | 10.6 |
| 1MNN | | NDT80 | 1.40 | R177 (A177) | WTT**TG**YGNc | 1.97 | -1.22 | 0.51 | 3.15 | -3.69 | 3.32 | 34.7 | 8.6 |
| 1CF7 | Winged-helix motif | E2F4-**DP2** | 2.60 | R121 (B121) | TTT**CG**CGCG | 2.87 | -1.95 | 1.05 | 3.66 | -2.61 | -4.35 | 34.9 | 13.1 |
| 1CF7 | | **E2F4**-DP2 | 2.60 | R56 (A56) | TTTCGCG**CG** | 2.52 | -1.65 | 0.27 | 3.43 | -6.21 | -10.1 | 35.1 | 17.2 |
| 1ECR | Interdomain Beta-strands | TUS | 2.70 | R232 (A232) | TAGTA**TG**TTGTAACTA | 2.75 | -1.11 | -0.86 | 4.27 | -0.49 | 9.84 | 35.9 | 18.7 |
| 1AAY | Zn Finger | ZIF268 | 1.60 | R18 (A118) | GCGTGGG**CG** | 2.23 | -1.14 | 0.31 | 3.37 | -4.5 | 6.05 | 33.7 | 2.3 |
| 1AAY | | | 1.60 | R74 (A174) | G**CG**TGGGCG | 2.42 | -0.95 | -0.21 | 3.49 | -7.23 | 7.82 | 36.2 | 4.8 |
| 1GU4 | bZIP | C/EBP beta | 1.80 | R289 (A289) | RTTR**CG**CAAY | 2.32 | -1.13 | -0.67 | 3.27 | 1.58 | 7.89 | 29.0 | 9.6 |
| | | | | | | AVG | -1.02 | -0.02 | 3.44 | -2.74 | 3.77 | 35.1 | 8.7 |
| | | | | | | STDEV | 0.57 | 0.56 | 0.31 | 2.61 | 6.36 | 2.8 | 6 |

[1] The first value is literature numbering and in parenthesis are the chain and residue number of the pdb file

[2] The distance between the centroid of the shifted pyrimidine and the centroid of the reference DNA pyrimidine

[3] As derived from values calculated by Olson et al. (1998)

* All distances and resolutions are in Angstroms

**Table 3.2 Summary of database search and DNA parameters for YpG steps contacted by histidines**

| PDB ID | Fold Type | Complex | Resolution | His residue | DNA consensus | Y to ref Y distance[2] | Shift | Slide | Rise | Tilt | Roll | Twist | Energy[3] (KT/2) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1AAY | Zn Finger | ZIF268 | 1.60 | H49 (A149) | GCG**TG**GGCG | 2.17 | -0.94 | -0.11 | 3.31 | -5.55 | 3.48 | 34.1 | 4.8 |
| 1PDN | Homeo-domain | PRD | 2.50 | H47 (C47) | **CG**TCACGSTTSR | 2.07 | -1.1 | 0.38 | 3.49 | -8.25 | 5.08 | 38.8 | 4.4 |
| | | | | | | AVG | -1.02 | 0.14 | 3.4 | -6.9 | 4.28 | 36.5 | 4.59 |
| | | | | | | STDEV | 0.11 | 0.35 | 0.13 | 1.91 | 1.13 | 3.28 | 0.26 |
| | | | | | Average CG/CG[3] | | 0.00 | 0.41 | 3.39 | 0.00 | 5.40 | 36.10 | |
| | | | | | Average TG/CA[3] | | -0.09 | 0.53 | 3.33 | -0.50 | 4.70 | 37.30 | |
| | | | | | ref TG/CA (1D98) | | 0.32 | 0.85 | 3.01 | -3.41 | 9.73 | 34.19 | |

[1] The first value is literature numbering and in parenthesis are the chain and residue number of the pdb file

[2] The distance between the centroid of the shifted pyrimidine and the centroid of the reference DNA pyrimidine

[3] As derived from values calculated by Olson et al. (1998)

* All distances and resolutions are in Angstroms

18.7 in terms of $k_BT/2$, and correspond to Z scores of 1.5 to 4.3 (Table 3.1, 3.2). These values indicate that the conformations of each 5'-YpG-3' step differ significantly from the average 5'-YpG-3' structures, and provide additional support for the idea that the conformations of these base pair steps have been deformed through interactions with the amino acid side chain.

**Arginine – 5'-YpG-3' interactions and sequence-specific recognition**

We next analyzed the available biochemical data to determine whether these proteins selectively bind to DNA targets that have pyrimidine rather than purine residues immediately 5' to the guanine base. In all cases, the available data strongly suggest that the 5'-pyrimidine is preferred and, in most cases, an analysis of the protein-DNA interface indicates that contacts between the arginine and the 5'-YpG-3' play a key role in this recognition.

For the AML/Runt protein, the consensus DNA binding site has been defined as 5'-YG**YG**GTY-3' (contacted bases in bold) through *in vitro* selection experiments (Kamachi et al., 1990; Melnikova et al., 1993; Speck and Terryl, 1995) , where the 5'-YpG-3' highlighted in bold is contacted by Arg174. In this case, no other contacts are made to the 5'-pyrimidine (or its complementary purine), although contacts are made to the phosphodiester backbone. We also note that the 5'-YpG-3' step immediately 5' to this step does not display significant unstacking, yet the 5'-pyrimidine is conserved.

For the homeodomain protein, MAT$\alpha$2, a 5'-**TG**T-3' sequence forms the core of the recognition site (Goutte and Johnson, 1993; Goutte and

Johnson, 1994) and Arg54 contacts the central guanine and stacks with the 5'-thymine (Li et al., 1998; Li et al., 1995). The O4 of the 5'-thymine hydrogen bonds with a water molecule that in turn is hydrogen bonded by Ser50. The Ser50 interaction is not conserved in all the MATα2 structures; for example, in the structure of MATα2 determined in the absence of MATa1 (**1APL**) (Wolberger et al., 1991), this water is missing and the T-A base pair in question makes no direct or indirect contact with the protein, other than through the DNA backbone. It therefore seems very likely that the strong preference for thymine at the 5' position is due to stacking interactions with Arg54. The homeodomain binding partner of MATα2, MATa1, also shows this form of recognition between Arg55 and a 5'-TpG-3' step within its consensus binding site, 5'-**TG**ATGA-3'. In this case, the 5'-thymine is not in contact with MATa1 other than Arg55, although the degree of displacement of the 5'-thymine is the smallest of the structures examined here. Another family of heterodimeric homeodomain transcription factors, Hox-Pbx (human) (Piper et al., 1999) and Ubx-Exd (drosophila) (Passner et al., 1999), show this kind of DNA interaction between the Pbx/Exd subunit the consensus DNA binding site 5'-A**TG**ATT-3' (Chang et al., 1996). Here, conserved Arg55 contacts the 5'-TpG-3' in bold via the major groove, while the A-T base pairs at the 5'-end of the site are also contacted by Arg5 via the minor groove. The minor groove contact by Arg5 is expected to exclude G-C base pairs at this position, but cannot select T-A over A-T base pairs. This selection is more likely provided by the Arg55 contact. This kind of interaction is reminiscent of the recognition

of the 5'-TpG-3' step at positions 5 and 6 of the Ndt80 binding site, where the thymine base appears to be pulled into the major groove through stacking interactions with Arg177, and at the same time, pushed via the minor groove by Pro57.

The binding site for the *E. coli* purine repressor, PurR, contains a conserved 5'-YpG-3' sequence at positions 3 and 4 of its consensus binding site (Rolfes and Zalkin, 1988; Rolfes and Zalkin, 1990). This 5'-YpG-3' is only contacted by a single arginine, Arg26, and is highly unstacked in a number of independently determined crystal structures (Glasfeld et al., 1999; Schumacher et al., 1994; Schumacher et al., 1997). Moreover, it seems likely that this kind of recognition is conserved within the large lacI family of transcriptional repressors, as many of these proteins recognize a conserved 5'-YpG-3' step at positions 3 and 4 which they contact with conserved arginine side chains(Rolfes and Zalkin, 1988).

The *E. coli* CRE recombinase also contacts its DNA targets using a helix-turn-helix motif (Guo et al., 1997; Guo et al., 1999). The recombinase recognizes a large, palindromic DNA target sequence, yet, paradoxically, makes only one major groove contact. The single contact is between Arg259 and a conserved 5'-CpG-3' step in the consensus sequence. Glu262 is in the vicinity of the 5'-cytosine, but its carboxylate group is too far from the cytosine exocyclic amine (> 4 Å) to exert a significant sequence selectivity.

The human E2F and related DP protein are transcription factors involved in cell cycle regulation that cooperatively recognize their target DNA

(5'-TTT**CG**CG**CG**-3') (Buck et al., 1995; Lees et al., 1993; Zhang and Chellappan, 1995) with a winged helix motif. In a heterodimeric structure of E2F-DP there are two 5'-YpG-3' steps that show arginine mediated unstacking (Zheng et al., 1999). One involves Arg56 of E2F and the second involves Arg121 of DP, both of which recognize 5'-CpG-3' steps. The step recognized by Arg121 of DP is simultaneously contacted from the minor groove by Arg17 of E2F. The major difference in this case is the arginine in the minor groove comes from a different subunit, E2F. The ability of these two subunits to cooperate in the recognition the 5'-CpG-3' step is dependent on the flexibility of the 5'-YpR-3' step to accommodate both contacts simultaneously.

ZIF268 contains three tandem Zn fingers that each recognize a 3 base pair DNA target within a 9 base pair site (Elrod-Erickson et al., 1996; Pavletich and Pabo, 1991). The N-terminal finger contacts the 3'-most 3 base pairs of the binding site (5'-G**CG**-3'). Arg18 immediately N-terminal to the recognition helix (at the "-1" position) contacts the 5'-CpG-3' step via hydrogen bonding and stacking. However, while the 5'-cytosine is clearly recognized in a sequence specific manner, at least part of this recognition is from major groove van der Waals interactions between the cytosine and Glu21. It seems plausible that in this case Arg18 and Glu21 cooperate to recognize the 5'-CpG-3' step, as the shift of the cytosine base into the major groove (facilitated by stacking with Arg18) is required to achieve van der Waals contacts with Glu21. In addition, the third Zn finger, which recognizes

the 5'-most binding site (5'-G**CG**-3'), contacts the highlighted guanine through Arg74. Here, the 5' C-G base pair does not make any other direct sequence specific contacts and the cytosine of this pair is shifted into the major groove. Therefore, specificity for the 5'- cytosine is likely achieved through arginine stacking.

The C/EBPβ homodimeric transcription factor preferentially binds a near palindromic DNA consensus sequence, 5'-RTTR**CG**CAAY-3' (Osada et al., 1996). The centre of symmetry of the DNA site is a 5'-CpG-3' step which is recognized in a surprisingly asymmetric manner by the protein. Arg289 from one of the monomers recognizes this base pair step through hydrogen bonding interactions with the guanine and van der Waals interactions with the 5'-cytosine, which induces unstacking between the contacted bases. In contrast, the 5'-CpG-3' on the complementary strand remains stacked and, as a result, Arg289 from the other monomer from the complex does not contact the complimentary strand but instead adopts a different conformation to contact an adjacent 5'-TpG-3' step.

The *E. coli* replication termination protein Tus recognizes DNA through a novel β-strand motif that interacts with a distorted DNA major groove (Kamada et al., 1996). A highly conserved 5'-TpG-3' step (Hill et al., 1988) is recognized by Arg232 from the β-strand motif. No other amino acid side chains contact this base pair step from either the major or minor groove. 5'-TpG-3' unstacking induced by Arg232 may also be involved in generating the DNA distortion that is recognized by Tus.

**Evidence for arginine-induced distortion 5'-YpG-3' steps**

The X-ray crystal structure of a mutant form of the MATα2 homeodomain bound to DNA provides strong evidence that arginine recognition of a 5'-TpG-3' step directly leads to the unstacking of the dinucleotide step. In the mutated protein, the arginine (Arg54) which normally contacts the 5'-TpG-3' step in the wild type protein, is mutated to alanine, as are two other major groove recognition residues, Ser50 and Asn51 (Ke et al., 2002). The DNA sequence of this structure is identical to the MATa1/α2/DNA structure described above. The wild type and mutant protein structures are very similar as indicated by an RMSD of 0.48 Å over 58 aligned Cα atoms. However, when the GC base pair of the mutant structure is aligned with the native structure, the extent of the unstacking induced by the arginine becomes apparent. The 5'-thymine of the wild type MATα2 is shifted more than 1 Å into the major groove when compared with the triple alanine mutant structure (Figure 3.4). Since neither Ser50 nor Asn51 make direct contact to the shifted base pair in the wild type structure, it is most likely that the displacement of the 5'-thymine into the major groove in the wild type structure is a direct consequence of its interactions with Arg54.

**5'-TpG-3' vs. 5'-CpG-3' recognition**

In theory, either 5'-TpG-3' or 5'-CpG-3' steps could be recognized equally well by simultaneous cation-π and hydrogen bonding interactions with

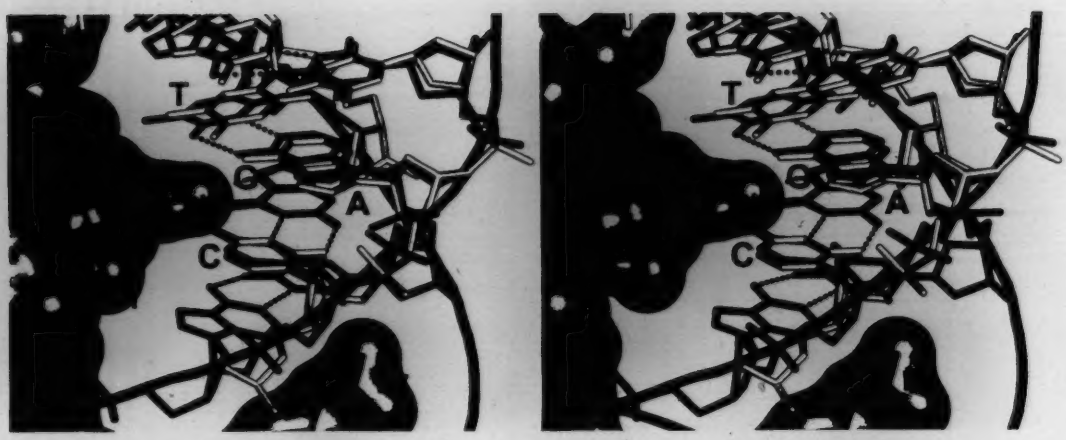


**Figure 3.4 Evidence for arginine-induced unstacking 5'-YpG-3' steps.** A triple alanine mutant of MATα2 (PDB ID: **1LE8**) was aligned to a wildtype MATα2 structure (PDB ID: **1AKH**) using the G-C base pair of the 5'-YpG-3' step as in Figure 3.1. The surface and protein side chains are from the wild type structure and the wild type DNA is shown in green. The DNA from the alanine mutant is shown in gold. The 5'-thymine base moves just over 1 Å towards the minor groove when Arg54 is mutated to an alanine.

arginine residues. In most cases, however, the proteins are specific for either a 5'-thymine or a 5'-cytosine. This specificity is often achieved by other elements of the protein. For example, in ZIF268-, Cre-, E2F- and DP-DNA complexes, substitution of a 5'-thymine for the consensus cytosine would result in a steric clash between the 5-methyl group of the thymine and a side chain in the major groove. For ZIF268 and Cre, glutamic acid residues provide this additional specificity, while tyrosine residues provide this function in E2F and DP. For the homeodomain proteins Pbx and Ubx, specificity for thymines is achieved through the minor groove contacts that would exclude the 2-amino group of guanine. In Ndt80 and MAT$\alpha$2, the specificity for thymines is in part accomplished with the aliphatic portion of the arginine side chain, which forms a hydrophobic half-pocket for the 5-methyl group of the thymine. For Ndt80, the importance of the 5-methyl group has been directly demonstrated at the 6$^{th}$ position of the consensus DNA target sequence. Substitution of the conserved thymine at this position with uracil, effectively replacing the 5-methyl group with a hydrogen atom, results in a 2-fold reduction in binding affinity (Lamoureux et al., 2002). However, there are some cases (Ndt80, AML1 and PurR, see Table 3.1) where the consensus sequence indicates that either pyrimidine can be accommodated. These examples show that recognition of 5'-YpG-3' steps by arginine has the flexibility to allow either pyrimidine in the 5' position.

**Histidine – 5'-YpG-3' recognition**

The planar, aromatic nature of histidine, together with its ability to hydrogen bond to nucleic acid bases, suggested that this side chain might also recognize 5'-YpR-3' steps by hydrogen bonding to the 3'-purine and stacking with the 5'-pyrimidine. To test this idea, we searched the protein-DNA sequence database for 5'-YpR-3' steps in which the purine N7 is within hydrogen bonding distance to a histidine, and the 5'-pyrimidine is displaced into the major groove. Two examples of such an interaction were found (Table 3.2). One is within the ZIF268-DNA complex, where His49 of the central finger contacts the central guanine of the three base pair site recognized by this finger, 5'-TGG-3'. The 5'-thymine base is specifically recognized by the protein and this recognition was previously thought to involve stacking interactions with His49. We suggest that this stacking is made possible by the shift of the thymine into the major groove. The second example is found in the structure of the Paired (PRD) homeodomain protein bound to its consensus DNA (Xu et al., 1995). In this structure, His47 contacts the 5'-most 5'-CpG-3' step of the DNA consensus (**CG**TCACGSTTSR, where S is a guanine or cytosine) (Jun and Desplan, 1996). Neither the 5'-cytosine, nor its complementary guanine is contacted by the protein, other than by His47. This 5'-CpG-3' step is not, however, conserved in the binding sites for Pax proteins that are highly similar to the PRD protein.

**Generality of protein-induced unstacking of 5'-YpG-3'**

Here we have described a way in which arginine or histidine residues can recognize the inherent flexibility of 5'-YpG-3' dinucleotide steps. Might other amino acids also induce similar distortions in 5'-YpR-3' steps? Glutamine and asparagine side chains can recognize adenine bases via a pair of hydrogen bonds to the major groove face of the DNA, in a manner that is structurally similar to arginine-guanine recognition (Seeman et al., 1976). We searched the protein-DNA structure database for examples of glutamine or asparagine recognition of 5'-YpA-3' steps that induced unstacking between the 5'-pyrimidine and the adenine, and enhanced contact between the glutamine/asparagines and the pyrimidine. No such examples were found. Thus, it may be that distortion of the 5'-YpG-3' step requires the relatively strong cation-$\pi$ interactions between an arginine or histidine side chain and the shifted nucleic acid base (Rooman et al., 2002; Wintjens et al., 2000).

Unstacking (as defined in Figure 3.2) occurs in approximately 34% of all 5'-YpG-3' steps that are contacted by arginine residues. In contrast, only about 10 % of 5'-YpG-3' steps in the free DNA structure database display this kind of unstacking, indicating that arginine-induced 5'-YpG-3' unstacking may be a common but not universal consequence of these interactions. Nevertheless, we have shown that this mode of protein-DNA recognition is utilized by many of the major classes of transcription factors and in most cases helps to explain hitherto inexplicable protein specificity for its consensus DNA. Our analysis has focused on relatively dramatic examples of

unstacking, however, it is possible that smaller distortions could also allow recognition of 5'-YpG-3' steps by these side chains. If this is the case, then 5'-YpG-3' recognition may be a much more general phenomena than reported here.

## REFERENCES


Bertrand, H., Ha-Duong, T., Fermandjian, S., and Hartmann, B. (1998). Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine-purine steps. Nucleic Acids Res *26*, 1261-1267.

Bravo, J., Li, Z., Speck, N. A., and Warren, A. J. (2001). The leukemia-associated AML1 (Runx1)--CBF beta complex functions as a DNA-induced molecular clamp. Nat Struct Biol *8*, 371-378.

Buck, V., Allen, K. E., Sorensen, T., Bybee, A., Hijmans, E. M., Voorhoeve, P. M., Bernards, R., and La Thangue, N. B. (1995). Molecular and functional characterisation of E2F-5, a new member of the E2F family. Oncogene *11*, 31-38.

Chang, C. P., Brocchieri, L., Shen, W. F., Largman, C., and Cleary, M. L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol *16*, 1734-1745.

Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., and Kennnard, O. (1989). Definitions and nomenclature of nucleic acid structure parameters. J Mol Biol *205*, 787-791.

Dickerson, R. E., and Chiu, T. K. (1997). Helix bending as a factor in protein/DNA recognition. Biopolymers *44*, 361-403.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure *4*, 1171-1180.

Fingerman, I. M., Sutphen, K., Montano, S. P., Georgiadis, M. M., and Vershon, A. K. (2004). Characterization of critical interactions between Ndt80 and MSE DNA defining a novel family of Ig-fold transcription factors. Nucleic Acids Res *32*, 2947-2956.

Glasfeld, A., Koehler, A. N., Schumacher, M. A., and Brennan, R. G. (1999). The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. J Mol Biol *291*, 347-361.

Goutte, C., and Johnson, A. D. (1993). Yeast a1 and alpha 2 homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. J Mol Biol *233*, 359-371.

Goutte, C., and Johnson, A. D. (1994). Recognition of a DNA operator by a dimer composed of two different homeodomain proteins. Embo J *13*, 1434-1442.

Guo, F., Gopaul, D. N., and van Duyne, G. D. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature *389*, 40-46.

Guo, F., Gopaul, D. N., and Van Duyne, G. D. (1999). Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. Proc Natl Acad Sci U S A *96*, 7143-7148.

Hepworth, S. R., Ebisuzaki, L. K., and Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in Saccharomyces cerevisiae. Mol Cell Biol *15*, 3934-3944.

Hill, T. M., Pelletier, A. J., Tecklenburg, M. L., and Kuempel, P. L. (1988). Identification of the DNA sequence from the E. coli terminus region that halts replication forks. Cell *55*, 459-466.

Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A *47 ( Pt 2)*, 110-119.

Jun, S., and Desplan, C. (1996). Cooperative interactions between paired domain and homeodomain. Development *122*, 2639-2650.

Kamachi, Y., Ogawa, E., Asano, M., Ishida, S., Murakami, Y., Satake, M., Ito, Y., and Shigesada, K. (1990). Purification of a mouse nuclear factor that binds to both the A and B cores of the polyomavirus enhancer. J Virol *64*, 4808-4819.

Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N., and Morikawa, K. (1996). Structure of a replication-terminator protein complexed with DNA. Nature *383*, 598-603.

Ke, A., Mathias, J. R., Vershon, A. K., and Wolberger, C. (2002). Structural and thermodynamic characterization of the DNA binding properties of a triple alanine mutant of MATalpha2. Structure (Camb) *10*, 961-971.

Lamoureux, J. S., Maynes, J. T., and Glover, J. N. M. (2004). Recognition of 5'-YpG-3' sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. J Mol Biol *335*, 399-408.

Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C., and Glover, J. N. M. (2002). Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. Embo J *21*, 5721-5732.

Lees, J. A., Saito, M., Vidal, M., Valentine, M., Look, T., Harlow, E., Dyson, N., and Helin, K. (1993). The retinoblastoma protein binds to a family of E2F transcription factors. Mol Cell Biol *13*, 7813-7825.

Li, T., Jin, Y., Vershon, A. K., and Wolberger, C. (1998). Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. Nucleic Acids Res *26*, 5707-5718.

Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science *270*, 262-269.

Lu, X. J., Shakked, Z., and Olson, W. K. (2000). A-form conformational motifs in ligand-bound DNA structures. J Mol Biol *300*, 819-840.

Melnikova, I. N., Crute, B. E., Wang, S., and Speck, N. A. (1993). Sequence specificity of the core-binding factor. J Virol *67*, 2408-2411.

Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., and Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A *95*, 11163-11168.

Osada, S., Yamamoto, H., Nishihara, T., and Imagawa, M. (1996). DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. J Biol Chem *271*, 3891-3896.

Ozsarac, N., Straffon, M. J., Dalton, H. E., and Dawes, I. W. (1997). Regulation of gene expression during meiosis in Saccharomyces cerevisiae: SPR3 is controlled by both ABFI and a new sporulation control element. Mol Cell Biol *17*, 1152-1159.

Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S., and Aggarwal, A. K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature *397*, 714-719.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell *96*, 587-597.

Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L., and Dickerson, R. E. (1987). Helix geometry, hydration, and G.A mismatch in a B-DNA decamer. Science *238*, 498-504.

Rolfes, R. J., and Zalkin, H. (1988). Escherichia coli gene purR encoding a repressor protein for purine nucleotide synthesis. Cloning, nucleotide sequence, and interaction with the purF operator. J Biol Chem *263*, 19653-19661.

Rolfes, R. J., and Zalkin, H. (1990). Autoregulation of Escherichia coli purR requires two control sites downstream of the promoter. J Bacteriol *172*, 5758-5766.

Rooman, M., Lievin, J., Buisine, E., and Wintjens, R. (2002). Cation-pi/H-bond stair motifs at protein-DNA interfaces. J Mol Biol *319*, 67-76.

Schumacher, M. A., Choi, K. Y., Zalkin, H., and Brennan, R. G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science *266*, 763-770.

Schumacher, M. A., Glasfeld, A., Zalkin, H., and Brennan, R. G. (1997). The X-ray structure of the PurR-guanine-purF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity. J Biol Chem 272, 22648-22653.

Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A *73*, 804-808.

Speck, N. A., and Terryl, S. (1995). A new transcription factor family associated with human leukemias. Crit Rev Eukaryot Gene Expr *5*, 337-364.

Tahirov, T. H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., *et al.* (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell *104*, 755-767.

Wintjens, R., Lievin, J., Rooman, M., and Buisine, E. (2000). Contribution of cation-pi interactions to the stability of protein-DNA complexes. J Mol Biol *302*, 395-410.

Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., and Pabo, C. O. (1991). Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67*, 517-528.

Xu, W., Rould, M. A., Jun, S., Desplan, C., and Pabo, C. O. (1995). Crystal structure of a paired domain-DNA complex at 2.5 A resolution reveals structural basis for Pax developmental mutations. Cell *80*, 639-650.

Zhang, Y., and Chellappan, S. P. (1995). Cloning and characterization of human DP2, a novel dimerization partner of E2F. Oncogene *10*, 2085-2093.

Zheng, N., Fraenkel, E., Pabo, C. O., and Pavletich, N. P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev *13*, 666-674.

# Chapter 4:

# Principles of protein-DNA recognition revealed in the structural analysis of Ndt80-MSE DNA complexes

**SUMMARY**

In this chapter I describe the X-ray crystal structures of Ndt80 bound to 10 distinct MSE variants. Comparisons of these structures with the structure of Ndt80 bound to a consensus MSE reveals structural principles that determine the DNA binding specificity of this transcription factor. The 5' GC-rich end of the MSE contains distinct 5'-YpG-3' steps that are recognized by arginine side chains through a combination of hydrogen bonding and cation-$\pi$ interactions. The 3' AT-rich region is recognized via minor groove contacts that sterically exclude the N2 atom of GC base pairs. The conformation of the AT-rich region is fixed by interactions with the protein that favor recognition of poly(A)-poly(T) versus mixed AT sequences through an avoidance of major groove steric clashes at 5'-ApT-3' steps. The majority of this work will be published in Structure (reference to come)

## INTRODUCTION

The Ndt80-MSE structure has been refined to 1.40 Å, the highest resolution yet achieved for a transcription factor – DNA complex. The high resolution structures (Lamoureux et al., 2002; Montano et al., 2002), together with studies of the effects of DNA and protein mutations on binding affinity (Fingerman et al., 2004; Lamoureux et al., 2002; Montano et al., 2002; Pierce et al., 2003), have shed light on how Ndt80 specifically recognizes the MSE. The MSE has been well defined based on statistical analyses of the promoters of middle genes (Chu et al., 1998; Chu and Herskowitz, 1998; Wang et al., 2005) and by mutational analysis (Lamoureux et al., 2002; Pierce et al., 2003). In addition the new mode of 5'-YpG-3' recognition described in the previous chapter and the competition for MSE binding sites with Sum1 make Ndt80 - MSE - Sum1 an ideal model system for a "simple" transcriptional network. In fact a computational study by (Wang et al., 2005) proposed that the two qualities of positive autofeedback (Ndt80's autoinduction of its own expression) and competition for a single binding site (mutually exclusive binding of Ndt80 or Sum1 on a single MSE) may be required for the sharp temporal/spatial expression profiles that are seen in many developmental pathways. If this is the case then the Ndt80 - MSE - Sum1 system could become a simple model system to help understand how these more complicated transcription networks work in higher organisms.

Many natural MSEs contain single base-pair variations, which affect Ndt80 binding affinities and may be important in tuning their activities during

sporulation. Here we report the high resolution crystal structures of several Ndt80-DNA complexes containing MSE sequences with deviations from the consensus sequence that was used in the initial crystallization study (Table 4.1). Overall, the structures reveal that Ndt80 holds the DNA in a rigid conformation that is relatively resistant to structural variations in response to these changes. Analysis of these structures is consistent with a dominant role for recognition of the YpG steps in the 5' part of the MSE, and provides an explanation for the selective recognition of poly(A)-poly(T) at the 3' end over mixed AT sequences.

## EXPERIMENTAL PROCEDURES

Protein Expression and Purification and DNA purification are as described in Chapter 2.

### Crystallization and data collection

Ndt80(1-340)-MSE complexes were prepared to a protein concentration ranging from 10-20 mg/ml and a ratio of protein to DNA of 1:1. Crystals were grown at room temperature (20°C) using the hanging drop method in conjunction with streak seeding with wild type crystals. Variant G1A/A9T was used in the initial crystallization trials of Ndt80-DNA complexes but did not yield crystals until it was streak seeded using wild type Ndt80-DNA complex crystals. In most cases once mutant crystals were obtained further optimization was done using the mutant crystals for seeds in the streak seeding procedure. The reservoir solution contained 25-35% PEG 400, 50 mM bis-tris-propane pH 7.0, 100 mM NaCl, 50 mM $CaCl_2$, and 1-5 mM DTT. 2 $\mu$l of complex was mixed with 2 $\mu$l of reservoir to form the drop. Drops were streaked either immediately or following a day of equilibration. Streaking was done using a horse hair dipped in streaking solution and rinsed twice in the reservoir. The streak seeding solution was prepared by diluting a drop that contained small crystals ~100 fold using the reservoir solution. Crystals grew to a maximum size of ~ 100-400 $\mu$m in 1-2 weeks and were harvested and

**Table 4.1**. Comparison of previously reported binding and activation data and the RMSD of the variant structures in comparison to wild type.

| Complex[a] | MSE sequence | Relative Kd[b] | Fold decrease in binding[c] | % Activation[d] | RMSD (aligned to WT in Å)[e] | |
|---|---|---|---|---|---|---|
| | | | | | all atoms | DNA |
| WT (1MNN) | GAC ACA AAA | 1 | - | 100 | - | - |
| vG1A/A9T (2EUV) | AAC ACA AAT | - | - | - | 0.65 | 0.42 (457) |
| vG1C (2ETW) | CAC ACA AAA | 3.3 | 3.6 | 20 | 0.63 | 0.38 (480) |
| vA4G (2EUX) | GAC GCA AAA | 3.1 | 5.9 | 20 | 0.52 | 0.30 (493) |
| mA4T (2EUW) | GAC TCA AAA | 5.4 | - | 19 | 0.49 | 0.28 (479) |
| mC5T (2EUZ) | GAC ATA AAA | 4.4 | 50 | 14 | 0.48 | 0.25 (490) |
| mA6T (2EVF) | GAC ACT AAA | 4.0 | 8.3 | 18 | 0.50 | 0.29 (484) |
| mA7T (2EVG) | GAC ACA TAA | 3.7 | 5.6 | 10 | 0.50 | 0.24 (483) |
| mA7G (2EVH) | GAC ACA GAA | - | 7.1 | 37 | 0.51 | 0.26 (501) |
| mA8T (2EVI) | GAC ACA ATA | 3.2 | 3.2 | 77 | 0.53 | 0.25 (482) |
| mA9C (2EVJ) | GAC ACA AAC | - | - | - | 0.53 | 0.32 (493) |

[a] The value in parenthesis is the PDB identifier

[b] Kd of Ndt80(1-340) for the indicated DNA substrate as referenced in (Lamoureux et al., 2002)

[c] Fold decrease of Ndt80 (1-409) bound in comparison to a wild type substrate (Pierce et al., 2003). The reference reported % bound values, whereas this column is = 100% divided by the % bound values for easier comparison to the relative Kd values

[d] % activation of the DNA sequence in a reporter assay as referenced in (Pierce et al., 2003)

[e] RMSD of the variant structure in comparison to wild type using aligned atoms with the align function of PyMol (DeLano, 2002). In brackets is the number of aligned atoms in the DNA column.

frozen in reservoir solution containing up to 10% glycerol where necessary. Data were collected at SBC-CAT (BL 19-ID) at the Advanced Photon Source and at the Advanced Light Source (BL 8.3.1). All crystals belong to space group $C222_1$ (a = 70 Å, b = 79 Å, c = 161 Å +/- 2%) with one complex in the asymmetric unit and are isomorphous with the wild type crystals (Table 4.2). All data from APS were processed with HKL2000 (Otwinowski and Minor, 1997) and all data from ALS were processed with mosflm and scala in the CCP4 program suite (Collaborative Computational Project, 1994).

**Structure determination, refinement, and analysis**

The variant and mutant complexes were built using the wild type Ndt80-MSE complex as the starting point. All waters were removed from the original pdb as well as the bases of mutated positions and the bases immediately adjacent to the mutated position(s). The remaining protein-DNA model was used to phase $2|F_o|-|F_c|$ and $|F_o|-|F_c|$ maps using the mutant diffraction amplitudes. Manual modeling was done with O (Jones et al., 1991) and alternatively PyMol (DeLano, 2002). Refinement and the addition of waters were carried out using REFMAC (Collaborative Computational Project, 1994) and protein geometry was analyzed with PROCHECK (Laskowski et al., 1993) . The final models of all of the mutants are similar to wild type containing residues 33-139, 146-286, 294-335 of Ndt80, most if not all of the DNA, and between 280-350 water molecules. The atomic coordinates of the Ndt80-MSE variants and mutants have been deposited in the Protein Data Bank [PDB ID: vG1A/A9T (2EUV),

vG1C (2ETW), mA4T (2EUW), vA4G (2EUX), mC5T (2EUZ), mA6T (2EVF), mA7T (2EVG), mA7G (2EVH), mA8T (2EVI), mA9C (2EVJ)]. Structures were aligned using the align function of PyMol with default settings, which includes 2 iterative cycles of outlier rejection. DNA helical parameters and torsion angles of the structures were calculated using 3DNA (Lu et al., 2000) and DNA conformational energies were calculated as described (Olson et al., 1998). The minor groove width was measured directly as the distance between phosphorus atoms on opposite strands staggered by three base-pairs

**Table 4.2.** Summary of X-ray experiments of variant and mutant Ndt80-MSE complexes

| Crystal Data | 1MNN | v1 | G1C | mA4T | vA4G | mC5T | mA6T | mA7T | mA7G | mA8T | mA9C |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 70.13 | 69.23 | 70.3 | 69.99 | 69.96 | 70.28 | 70.19 | 70.43 | 69.53 | 69.71 | 69.63 |
| b | 78.81 | 79.25 | 78.92 | 78.34 | 79.13 | 78.84 | 78.75 | 78.78 | 78.44 | 78.55 | 79.04 |
| c | 161.39 | 160.88 | 161.63 | 161.48 | 161.83 | 161.65 | 161.54 | 162.67 | 161.33 | 161.78 | 161.68 |
| Space group | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 | C2221 |
| **Data Collection** | | | | | | | | | | | |
| Wavelength (Å) | 0.980 | 1.072 | 1.072 | 1.072 | 1.072 | 1.072 | 1.072 | 1.072 | 1.009 | 1.009 | 1.009 |
| Resolution (Å) | 100-1.4 | 17-1.95 | 24-1.67 | 24-1.68 | 36-1.57 | 24-1.56 | 24-1.56 | 27-1.55 | 44-2.0 | 35-1.8 | 36-1.9 |
| Unique Reflections | 88 384 | 35074 | 52485 | 50924 | 62985 | 64128 | 63947 | 65883 | 30725 | 41465 | 35986 |
| Completeness | 99.8(97.8) | 99.4 (100) | 98.6(99.0) | 98.7 (99.5) | 94.6(72.2) | 96.0(86.9) | 98.7(93.4) | 98.5(90.8) | 95.8(81.3) | 99.9(100) | 99.9(100) |
| I/sigma | 35(3.0) | 9.6 (1.4) | 18.4(3.8) | 17.8(1.6) | 15.1(3.0) | 15.2(2.1) | 19.1(3.1) | 14.1(2.5) | 23.2(2.4) | 21.8(2.5) | 23.6(2.73) |
| Rsym | 0.051(0.493) | 0.045(0.51) | 0.043(0.176) | 0.067(0.45) | 0.051(0.225) | 0.051(0.341) | 0.045(0.239) | 0.053(0.283) | 0.067(0.510) | 0.090(0.806) | 0.074(0.612) |
| Redundancy | 38(5.2) | 4.1(3.8) | 4.7(2.7) | 8.9(5.4) | 4.3(1.9) | 4.6(3.1) | 5.6(3.2) | 4.9(3.0) | 3.4(2.3) | 7.5(7.5) | 3.8(3.7) |
| **Refinement** | | | | | | | | | | | |
| Rcrys/Rfree | 19.4/20.6 | 19.3/23.3 | 18.0/19.9 | 17.7/20.1 | 17.1/20.3 | 17.6/19.4 | 17.2/19.7 | 16.9/18.4 | 17.9/21.2 | 17.2/21.1 | 17.2/20.2 |
| RMSD bonds | 0.011 | 0.011 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.009 | 0.008 | 0.011 | 0.008 |
| RMSD angles | 1.52 | 1.45 | 1.38 | 1.39 | 1.45 | 1.42 | 1.41 | 1.40 | 1.33 | 1.45 | 1.27 |
| **Ramachandran** | | | | | | | | | | | |
| Favored/Allowed | 90.6/8.2 | 90.2/8.6 | 90.5/8.8 | 92.9/6.3 | 92.5/6.7 | 92.2/7.1 | 92.2/7.1 | 92.2/6.7 | 91.8/7.1 | 92.2/6.7 | 93.3/5.5 |
| Generously Allowed/Disallowed | 0.8/0.4 | 0.8/0.4 | 0.4/0.4 | 0.4/0.4 | 0.4/0.4 | 0.4/0.4 | 0.4/0.4 | 0.8/0.4 | 0.8/0.4 | 0.8/0.4 | 0.8/0.4 |

Values in parentheses refer to the highest resolution shell

## RESULTS

The variant and mutant MSE sequences used in this study are summarized in Table 4.1 and a summary of the X-ray experiments are presented in Table 4.2. In most cases, we attempted to crystallize complexes containing a single base-pair substitution which resulted in small but significant defects in binding affinity. The MSEs are classified into 2 groups: variants (specified with the prefix "v") in which the MSE sequence is changed from the original or 'wild type' sequence but retains the minimal MSE consensus requirements, and mutants (specified with the prefix "m"), in which one of the conserved base-pairs has been altered. In all cases the crystals contained the DNA-binding domain of the Ndt80 protein from residues 1-340 and a 14 mer DNA with a single 5' overhanging nucleotide. We attempted to crystallize complexes containing a total of 20 different mutant and variant MSEs. Of these, 9 yielded crystals which diffracted X-rays to high resolution (between 1.56 and 2.0 Å resolution). Of the 11 complexes that did not produce satisfactory crystals; 5 were single mutations at position 3 or 5 that were expected to significantly reduce the binding affinity, 3 were double mutations at semi- or non-conserved positions resulting in variant MSEs, and 3 were single mutations of the remaining conserved positions that were expected to marginally decrease the binding affinity.

All of the crystals obtained were isomorphous with the wild type, and have been refined to their respective resolution limits (see Experimental procedures, Table 4.2). The accuracy of these structures is sufficient to define not only the protein geometry, but also the DNA backbone torsion angles and the protein-

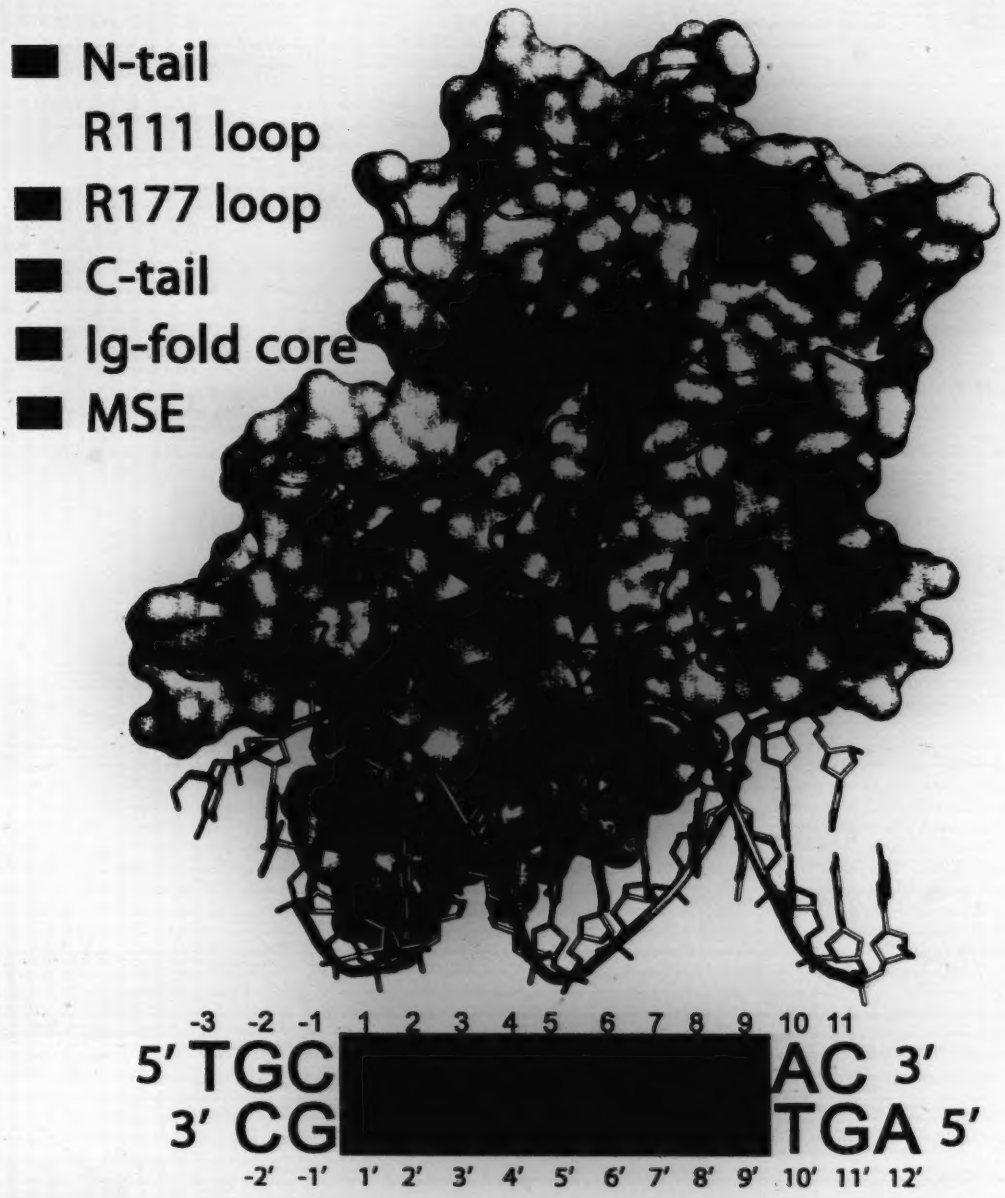


**Figure 4.1 Overall diagram of the Ndt80-DNA complex.** The key DNA recognition loops are highlighted. The DNA used in the wild type structure is shown below with the MSE highlighted and numbered.

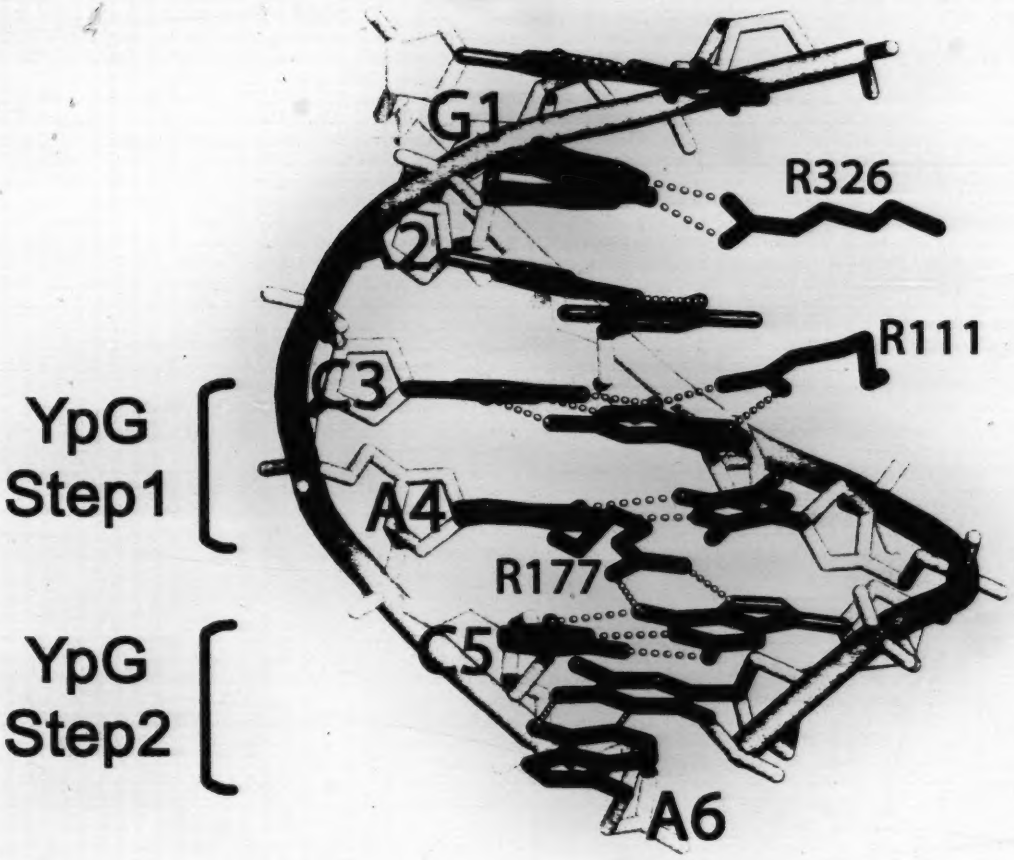


**Figure 4.2 Major groove view of the wild type structure.** Note the high degree of base unstacking of the 5' pyrimidines in both 5'-YpG-3' steps.

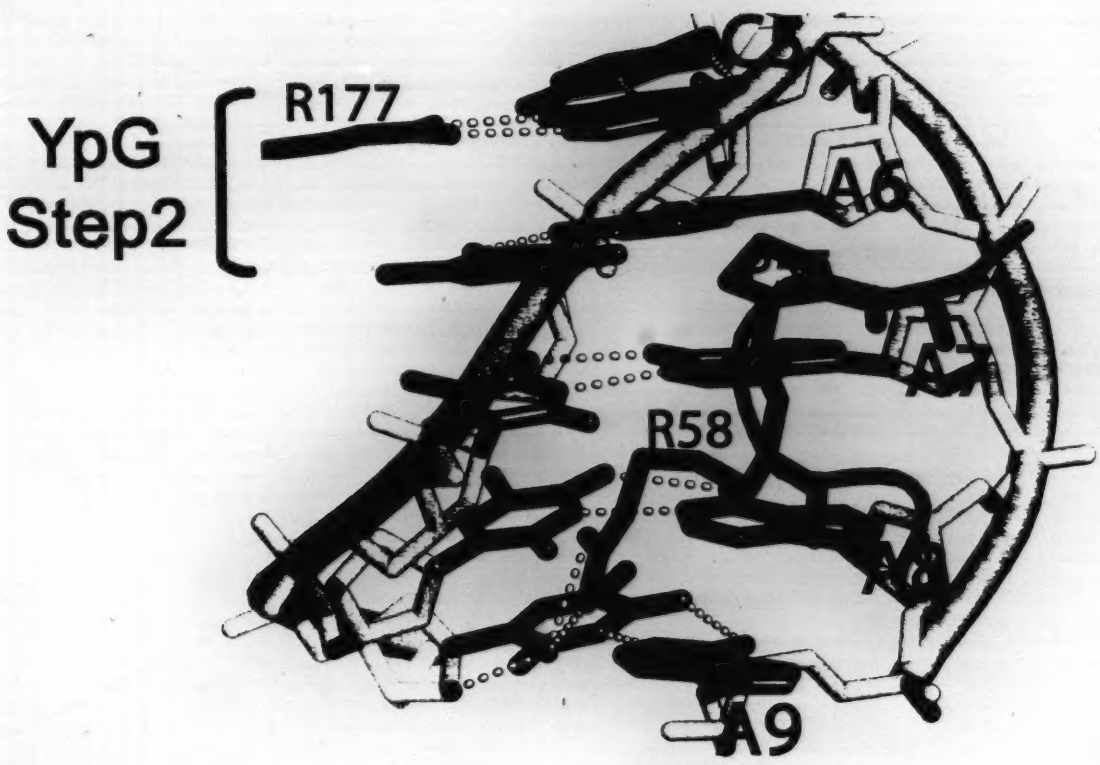


**Figure 4.3 Minor groove view of the wild type structure.** The steric constraints of residues Pro57 and Arg58 span four base pairs or the entire poly-A tract. The unstacking of the second YpG step is even more apparent in this view.

DNA contacts, as well as networks of water molecules trapped at the protein-DNA interface that mediate recognition.

**Variants**

**vG1A/A9T**

This was the only structure we obtained of a complex containing two substitutions. The substitutions are at semi-conserved positions at the edges of the MSE that do not seem to play a major role in binding affinity, but nevertheless may be important in fine-tuning the relative affinities of a particular MSE for either Ndt80 or Sum1. The overall structure is very similar to the wild type structure, with an RMSD for Cα atoms of 0.43 Å (over 290 atoms) (Table 4.1). The most striking change in this variant occurs at position 1 of the MSE where a guanine base is recognized by Arg326 in the wild type structure. In this variant the guanine is replaced with an adenine and the bidentate hydrogen bonding to the arginine cannot occur (Figure 4.4). Instead, the arginine side chain shifts to hydrogen bond with the backbone phosphate at the -1 position. The density in this position suggests that there is an alternate conformation of this arginine residue in which the Cδ-Nε angle is rotated 180°. In both conformations, there is a hydrogen bond to the carbonyl oxygen of Ser259 and weak hydrogen bonds to the phosphate. In response to the arginine reorientation, the DNA shifts such that the cytosine 5' to the substituted adenine slides back ~1 Å to stack over the

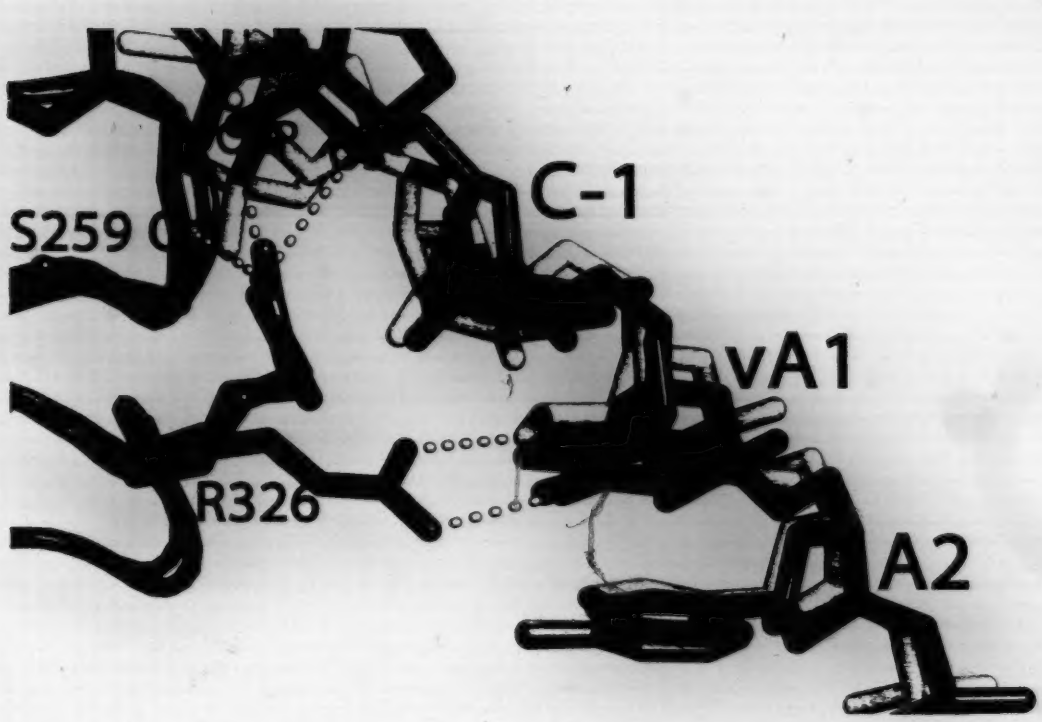


**Figure 4.4 Structural rearrangements seen in the vG1A/A9T complex**. There is a large shift of Arg326 residue which makes new contacts to the DNA backbone and also appears to have an alternate conformation. The variant is colored blue and purple, for DNA and protein respectively. The wild type structure is semi-transparent orange and teal for DNA and protein respectively.

adenine, and the 5' phosphate of the cytosine is repositioned to contact the arginine guanidinium. In the model this shift is accomplished by a change in the backbone ε and ζ torsion angles from the common BI conformation (although not a proper BI) where ε and ζ are in the ($t/g^-$) range and ε - ζ ≈–90 to the less frequently observed BII conformation where ε and ζ are in the ($g^-/t$) range and ε - ζ ≈+90 at the -1 position (Prive et al., 1987). Interestingly the DNA appears to be poorly ordered in this structure in comparison to the others. In fact, this variant was the only one in which the density was poor enough to exclude modeling the nucleotides at the -2', and -3 positions.

The other variation is at the opposite end of the MSE consensus where the position 9 adenine is changed to a thymine. This change maintains the consensus for an MSE sequence. In the wild type structure Arg58 is packed into the minor groove in a conformation in which its aliphatic portion of the sidechain (C$\beta$, C$\gamma$, and C$\delta$) mimics Pro57 (C$\beta$, C$\gamma$, and C$\delta$) and serves to exclude the N2 of a guanine from the minor groove in the poly-A tract of the MSE. The guanidino group of this arginine co-ordinates a water molecule that appears to serve a similar role to exclude G-C base-pairs at position 9 of the MSE. In this variant the position of the DNA backbone and base-pairs are extremely similar to the wild type structure. The only minor variation occurs with a slight shift of the coordinated water accommodating the change of its hydrogen bonding partner from a nitrogen to an oxygen.

**vG1C**

The changes seen with this variant are very similar to those observed for vG1A/A9T (RMSD (Cα atoms) = 0.42 Å). Arg326 cannot hydrogen bond the substituted cytosine at the +1 position and therefore swings out to contact the DNA backbone as is seen in vG1A/A9T (Figure 4.4). However there are 2 differences between these two variants. The first is this arginine does not appear to have alternate conformations. Secondly the shift of the DNA backbone is not accomplished with a switch from the BI to BII conformation but merely a displacement of the DNA with maintenance of torsion angles closer to the wild type structure. This variant also lacks interpretable density for the 5' overhanging thymine, however, in contrast to vG1A/A9T, the -2' base-pair can be modeled in this variant. A caveat for both vG1C and vG1A/A9T mutants is that, due to the relatively poor electron density around position 1, torsion angles can not be accurately determined; however the shift from BI to BII can be reasonably inferred as it involves a characteristic shift in the relative position of the O3' atom.

**vA4G**

The final complex of a variant MSE bound to Ndt80 was obtained by replacing the position 4 AT pair found in the wild type structure with a GC, maintaining the requirement for a purine at position 4. Remarkably, despite maintaining a consensus MSE sequence this variant had the largest localized DNA backbone shift seen in all of the complexes investigated in this study

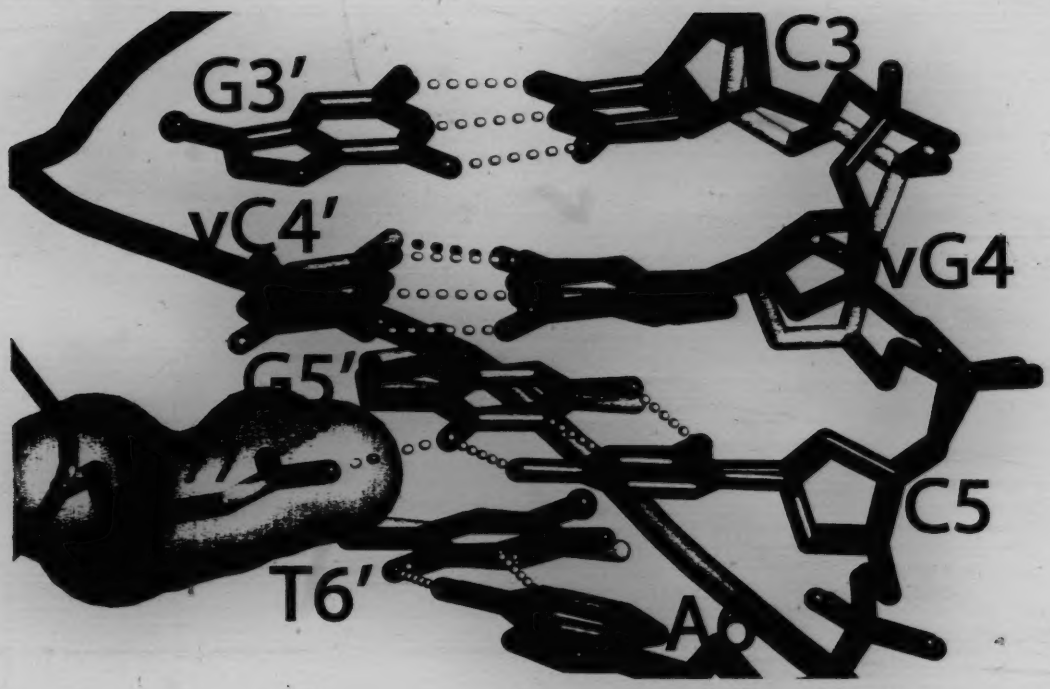


**Figure 4.5 Structural rearrangements seen in the vA4G complex**. The significant opening (~7°) of this base pair in the wild type structure does not pose a problem for an A-T base pair however when replaced with a G-C pair the guanine N2 and cytosine O2 distance would be too short (illustrated with a red hydrogen bond) if the bases are not rearranged. The variant is colored blue and purple, for DNA and protein respectively. The wild type structure is semi-transparent orange and teal for DNA and protein respectively.

(Figure 4.5). The driving force for this structural rearrangement probably lies in the fact that the TA pair in the wild type structure is significantly opened toward the major groove by about 7° relative to the mutant. While this opening does not dramatically distort the two hydrogen bonds of the AT pair, if one simply replaces the AT pair with a GC in this opened conformation, the guanine N2 – cytosine O2 distance is too short for a stable hydrogen bond (2.3 Å). The cytosine in the variant also shifts slightly (~0.3 Å) into the major groove, probably to increase stacking interactions with Arg111, but exerting further pressure on the guanine partner. As a result, the guanine of the opposite strand shifts out 1.4 Å into the major groove to maintain good hydrogen bonding geometry, facilitated by a shift to the BII conformation in this nucleotide. The fact that the guanine nucleotide is not contacted by the protein probably allows this movement and, as a result, there is only a small loss of binding affinity associated with this mutation.

**Mutants**

**mA4T and mA6T**

Each of these mutations destroys the MSE consensus by mutating the 5' pyrimidine in one of the two critical 5'-YpG-3' steps to an adenine, resulting in 5'-ApG-3' steps. What is most remarkable about these mutations is the lack of change in the structures (Figure 4.6). Despite disrupting the consensus MSE, the conformation of both the DNA and protein is almost entirely unaltered. Even the hydrogen bonding in the minor groove is maintained because the positions of hydrogen bond acceptors and donors do not change. Only in the major groove

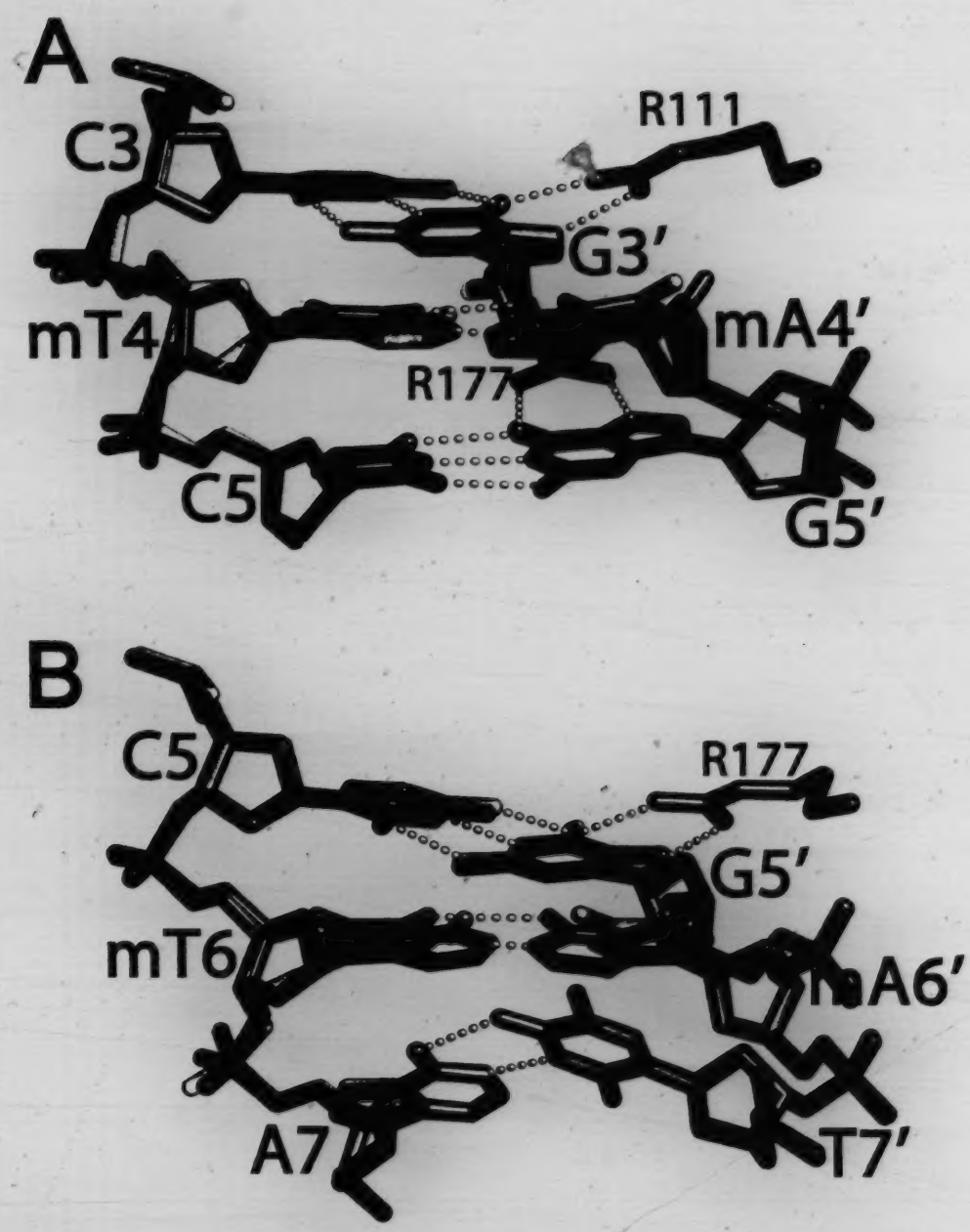


**Figure 4.6. Mutation of the pyrimidines in the 5'-YpG-3' steps**. In both panels the wild type structure is semitransparent and illustrated in orange and teal for DNA and protein respectively. The mutants are colored blue and purple for DNA and protein respectively. **(A)** mA4T complex. **(B)** mA6T complex. In both cases, despite the relatively large decrease in binding affinity, there are surprisingly few structural changes. The expected relaxation of unstacking when the YpG step is mutated was not observed; instead the DNA appears to be anchored in place.

does the hydration shell vary as would be expected with a change in the positions of the hydrogen bonding partners. The DNA backbone is held rigidly in place through extensive contacts with Ndt80, such that the substituted 5' adenosine is held in the BII conformation shifted into the major groove. In spite of the fact that the N9 of the adenine in the mutants, and the N1 of the corresponding thymines in the wild type structure, are in virtually the same position, the reduced size of the purine imidazole, compared to the thymine, results in significantly less van der Waals contacts between the adenine and the arginine.

The conformational energy of the dinucleotide steps in each of these structures can be estimated using helical parameters. The energy values are expressed in terms of $k_BT/2$, relative to the mean values for free B-DNA (Olson et al., 1998). For mA4T the energy of the mutated ApG step is 67 and for mA6T it is 64 whereas the corresponding TpG steps in the wild type structure are 44 and 27 respectively. It is important to note that these two TpG steps are already the highest energy steps in the wild type structure; indeed, the average conformational energy for dinucleotide steps in the wild type structure is 14. The reduced van der Waals contact, together with the additional cost of unstacking the more rigid dipurine step, largely explains the small but biologically significant loss of binding affinity (~2-3 – fold) for these sequences compared to the consensus MSE.

**mC5T**

Our previous studies demonstrated that the conserved CG base-pairs at

positions 3 and 5 are the most sensitive to mutation. Although we attempted to crystallize all possible substitutions at these positions, we were only able to crystallize mC5T. This mutation is the best tolerated of the three possible mutations at this position and results in a 3-fold decrease in affinity in comparison to the wild type MSE (Lamoureux et al., 2002). This structure has the lowest RMSD (all aligned atoms) of all the variants and mutants (RMSD = 0.48 Å). The reason mutations are poorly tolerated in this position is that the guanine at position 5' along with the thymine at position 6' comprise the central 5'-YpG-3' step of the MSE consensus. This step is recognized by Arg177 in the major groove and Pro57 in the minor groove as well as extensive backbone phosphate and sugar contacts (Figure 4.3). This region of the DNA is almost completely encompassed by Ndt80 and accounts for a great number of the protein-DNA contacts. The substitution of the CG by the TA base-pair at this position makes it impossible for Arg177 to make bidentate hydrogen bonds to the guanine base. Instead, Arg177 rotates ~180° about $\chi 2$, displacing 2 water molecules in the process (Figure 4.7). The two displaced waters are part of a cluster of 6 well ordered waters that mediate interactions between the protein and base-pairs 4 and 5. The guanidinium group of Arg177 in the C5T structure makes hydrogen bonds that replace some but not all of those made by the displaced water molecules. As a result of this weakened hydrogen bonding network, the Arg177 guanidinium group is not held in place as tightly as in the wild type structure, resulting in higher B factors for the guanidinium atoms (~30 $Å^2$) compared to the wild type structure (~18 $Å^2$).

**Figure 4.7. Stereo view of the changes in the mC5T structure**. The wild type structure is semitransparent and illustrated in orange and teal for DNA and protein respectively, waters are red and hydrogen bonds are yellow. The mutant structure is colored blue and purple for DNA and protein, respectively. The waters in the mutant are colored green and the hydrogen bonds are beige. The surface of the protein, excluding the side chain for Arg177, from the mutant structure is rendered in grey and is essentially identical to that of the wild type structure. The mutation of the guanine recognized by Arg177 to an adenine forces the side chain to reorient, displacing two water molecules.

**mA7T**

In the wild type structure the poly-A tract is contacted predominantly through the minor groove and the backbone and beyond position 6 of the MSE there are no direct side-chain contacts with the major groove at all (Figure 4.3). As mentioned above, G-C base-pairs are excluded in this region by the steric clashes that would occur in the minor groove by a 2 - amino group of a guanine base, but it is not clear why an A-T to T-A base-pair substitution is disfavored. This mutation introduces no new steric clashes with the protein, yet the affinity is reduced by approximately 3-5 fold (Figure 4.8) (Lamoureux et al., 2002; Pierce et al., 2003). The position of the backbone and bases at the substituted thymine at position 7 remains quite similar to that of the wild type structure despite a subtle shift from BI towards a more BII-like conformation (from $\epsilon$- $\zeta$= -28 in wild type to +12 in mA7T). At position 6 there is the opposite shift from the rarer BII to the typical BI conformation which repositions the adenosine base towards the minor groove by approximately 0.6 Å. Previous work has noted that the BII conformation is often associated with an unstacking of the base relative to the 3' base (Prive et al., 1987). The change from BII to BI at position 6 and the additional BII character at position 7 seen in this mutant serve to increase the distance between the adenosine at position 6 and the introduced mutant adenosine on the opposite strand at position 7, yet the A6(N6)-A7'(N6) distance is still 3.0 Å, indicating these two atoms clash despite the rearrangements of the DNA. This clash, in addition to the rearrangements that are required to minimize it, are quite likely responsible for the 3-5 fold decrease in affinity of Ndt80 for this substrate.

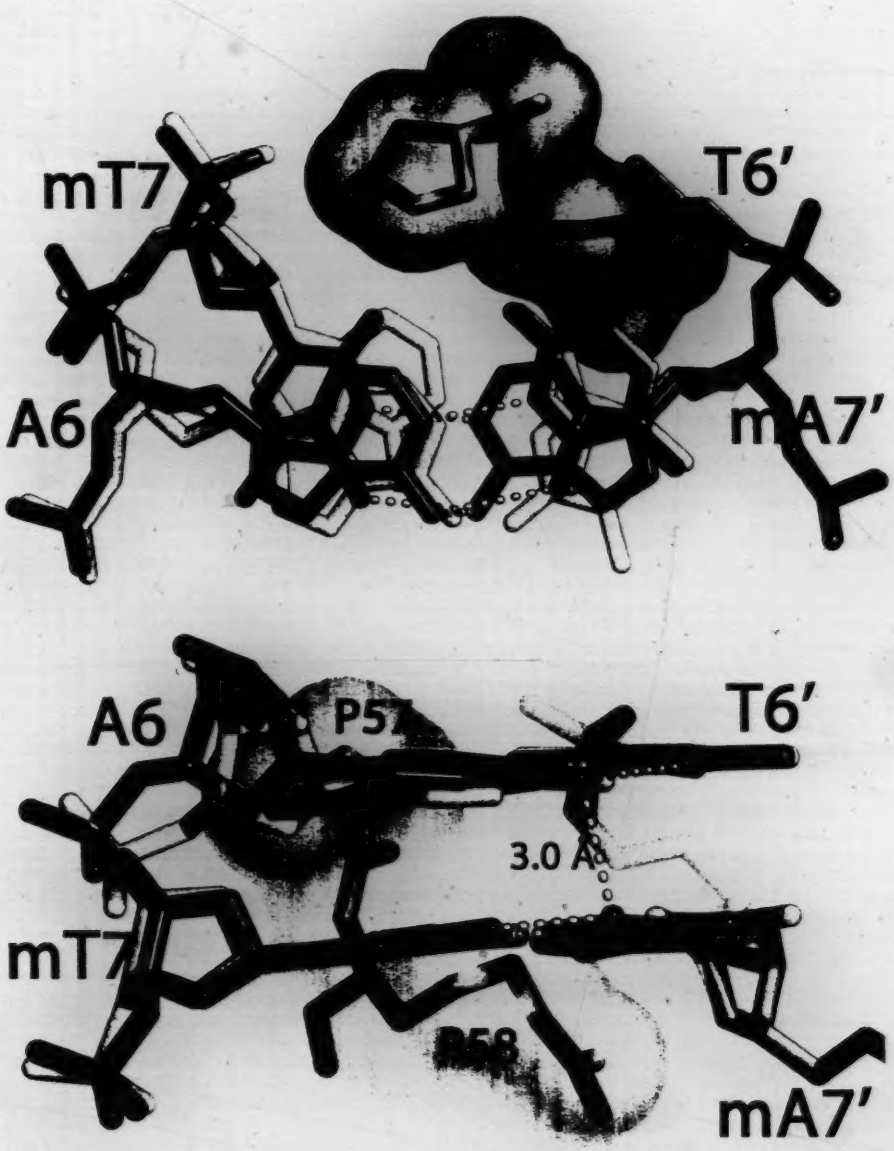


**Figure 4.8 Structural rearrangements of the mA7T mutant.** The mutant is colored blue and purple for DNA and protein respectively with the $\epsilon$ and $\zeta$ bonds that define the BI and BII conformations highlighted in cyan, superimposed on the DNA from the wild type structure (semi-transparent orange). The top portion of each panel is a view down the helical axis and the bottom portion is a perpendicular view from the major groove. The change from a BII to BI conformation at position 6 and the concurrent shift from BI to BII at position 7 maximize the cross strand N6 distance.

**mA8T**

This mutant is similar to mA7T both in the type of mutation and its context within the MSE (Figure 4.9). In addition similar structural changes are observed in this mutant although to a lesser degree. As is seen in the mA7T mutant, the base-pairs of mA8T are in nearly identical positions as the wild type structure. It is the adenosine at position 7 that, once again, shifts toward the minor groove by approximately 0.3Å. This shift is accomplished by several changes in the backbone torsion angles, but unlike mA7T, there is not a clear shift from BII to BI as the wild type structure is not in a BII conformation at this position. However, the adjustment at position 7 makes the mutant more "BI like" as the $\epsilon$-$\zeta$ 5' to the mutated adenine is -65° versus -28° seen in the wild type structure. The additional BI character is likely due to a steric clash that would otherwise occur between the 5-methyl group of the mutated thymine and its phosphate. The thymine paired with the shifted adenine at position 7, however, remains in an identical position to that seen in the wild type structure, due to the extensive backbone contacts with this strand of the DNA, effectively locking the DNA in place. Again it appears that these rearrangements serve to increase the distance between the adjacent adenines on opposite strands, resulting in an N6-N6 distance of 3.1 Å across the major groove.

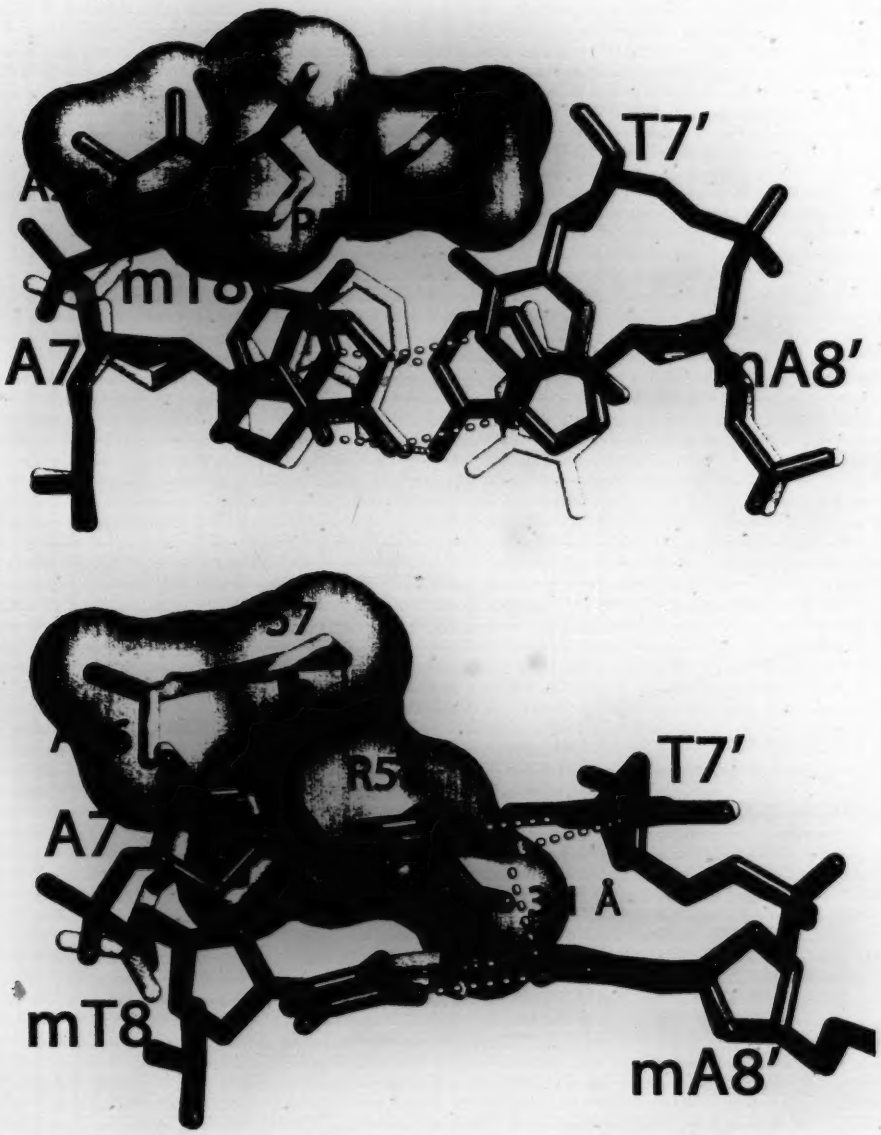


**Figure 4.9 Structural rearrangements of the mA8T mutant.** The mutant is colored blue and purple for DNA and protein respectively with the $\epsilon$ and $\zeta$ bonds that define the BI and BII conformations highlighted in cyan, superimposed on the DNA from the wild type structure (semi-transparent orange). The top portion of each panel is a view down the helical axis and the bottom portion is a perpendicular view from the major groove. This mutant sees similar changes as mA7T but to a lesser extent. There is also a potential clash between the position eight 5-methyl group and its phosphate oxygen.

**mA7G**

The structure of the wild type complex suggested that GC base-pairs would be disfavored at this position due to steric clash between the 2-amino group of the guanine base and Arg58 in the minor groove. The structure of the mutant reveals that while the substituted cytosine adopts the same conformation as the thymine in the wild type structure, the guanine base shifts to reduce its steric clash with Arg58 (Figure 4.10). However, the magnitude of this shift is modest, only about 0.4 Å toward the major groove. This does not completely relieve the steric clash, as the guanine N2 – Arg58 C$\gamma$ distance is still a rather short 3.2 Å, in contrast to the 4.1 Å adenine C2 - Arg58 C$\gamma$ distance observed in the wild type structure. A change from the BI to BII conformation at position 7 assists the shift of the guanine base toward the major groove and the concurrent change from BII to BI at position 6 serves to bring the rest of the DNA back into register with the wild type structure.

The protein does not seem have any noticeable changes, in particular Arg58 is in a nearly identical conformation as in the wild type structure. It is interesting that it does not change conformation to avoid the close proximity to the 2-amino group of the mutant guanine, as most of the volume around this side chain contains water molecules that one might think could be easily displaced. This is not the case and either the contacts made by this arginine hold it securely in place and/or the waters around this arginine are integral components of the structure. The B-factors of nearly all of these waters are < 30Å and their positions are conserved, consistent with the view that they are indeed critical to the structure and hence the recognition of the MSE.

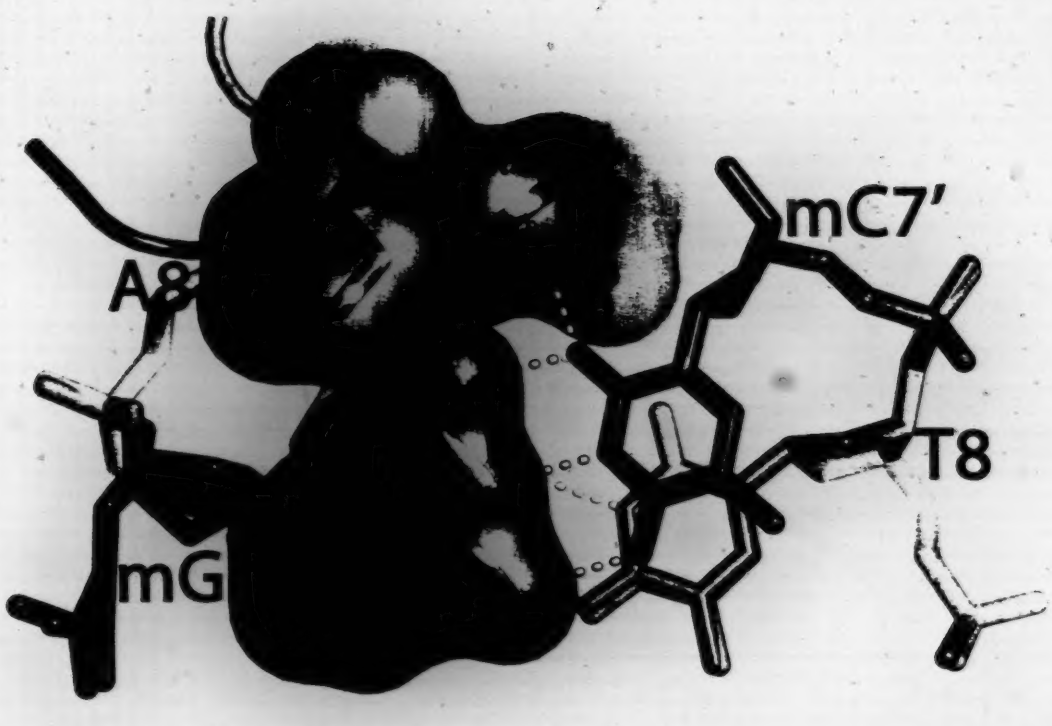


**Figure 4.10 Structural consequences of the mA7G mutation.** Color scheme follows the previous panels. Mutation to a guanine introduces a N2 amino group and consequently a clash with the Arg58 side chain.

**mA9C**

This mutant structure is interesting in that it provides a possible explanation for the specificity at this position that was lacking in the original paper (Figure 4.11). The only contact between this position and the protein is made through backbone contacts and water-mediated hydrogen bonds in the minor groove. In the original structure we did not see why G-C base-pairs were excluded from this position. It appeared that the 2-amino group of a guanine base could simply displace a water molecule in the hydrogen bonding network of the minor groove. This mutant structure demonstrates that the consequences of such a mutation may not be as trivial. In the wild type structure three of the waters that make up this hydrogen bonding network (W1, W2, and W3) are tightly hydrogen bonded to their maximum number of hydrogen bonding partners, 2 donors and 2 acceptors (Figure 4.11). Nearly all of these hydrogen bonds are relatively strong with the majority having a distance of less than 3.0 Å. The introduction of a guanine base displaces a single water molecule (W1) but the N2 amino group is too far away to contact the hydrogen bonding partners of this displaced water. In effect, this mutation eliminates all 4 hydrogen bonds of the water it displaced. The whole network of waters in the minor groove seems to become destabilized if one uses B-factors as an indicator of relative stability. Despite the fact that the average B-factor for this mutant is lower than that of the wild type structure, the remaining two waters in the mutant structure (W2 and W3) are both 10 $Å^3$ higher than that seen in the wild type structure, indicating that these waters are destabilized. Interestingly, Arg58 has essentially identical B-

factors in both the wild type and mutant structure indicating the difference seen in the B-factors of these waters is not simply a general localized disorder. If this explanation is indeed how Ndt80 recognizes position 9, then we would expect it to be an extremely weak preference for an A or T. In fact two recent papers call into question the importance of this position for defining an MSE (Pierce et al., 2003; Wang et al., 2005).

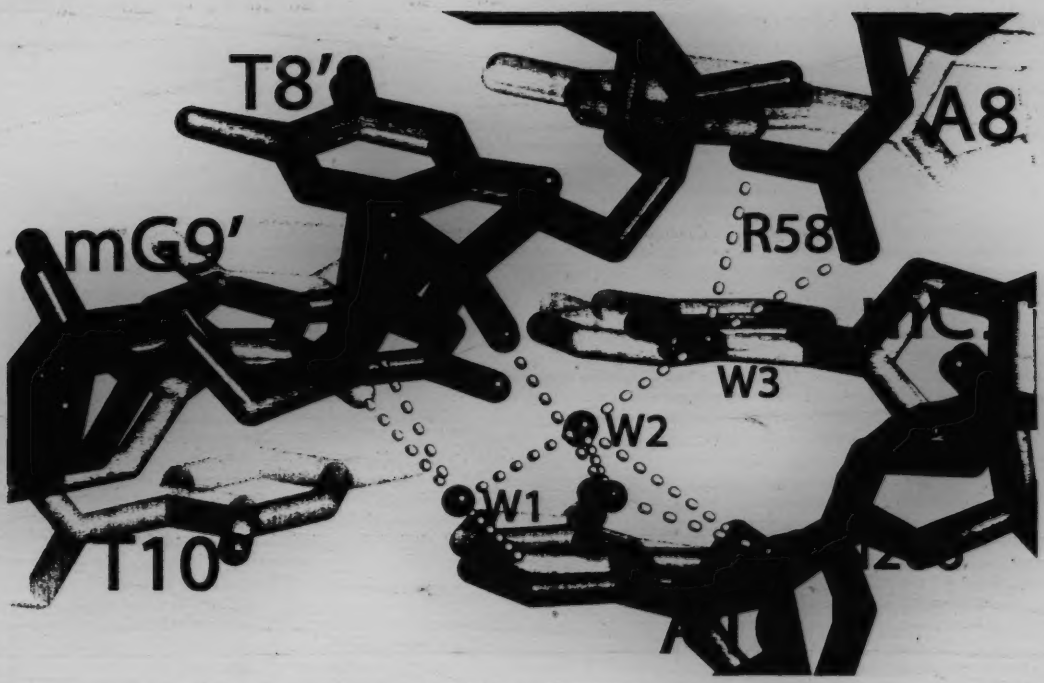


**Figure 4.11. Structural consequences of the mA9C mutation.** The red spheres are the waters found in the wild type structure while the green ones are those of the mutant structure. Hydrogen bonds in yellow are those found in both the wild type and mutant structures, while those in beige are wild type hydrogen bonds that are lost in the mutant.

## DISCUSSION

This large set of mutant and variant complex structures provides a basis to understand the fundamental and subtle mechanisms of recognition utilized by Ndt80 in order to bind to its appropriate MSE target sequences. There is good agreement between the structural data and the mutational studies both of the MSE and protein residues. Mutational studies of Ndt80 highlight the importance of Arg111, Arg177, Pro57, Arg58 in MSE recognition and affinity (Fingerman et al., 2004; Montano et al., 2002). In addition there are several residues that these studies highlight that cannot be directly rationalized through structure analysis alone. While the idea of a simple universal code that could be used to predict DNA-binding preferences of a given protein *a priori* seems unrealistic, perhaps the principles of recognition shown here might be applicable across transcription factor families and provide more insight into protein-DNA interactions as a whole. One of the most striking observations that can be made of all for these structures is their very high similarity to the wild type structure. The structure with the largest RMSD was vG1A/A9T with an RMSD of 0.66 Å over all aligned atoms (~2800 atoms) or 0.43 Å over the Cα atoms (289 atoms). This is a very close alignment when compared to the RMSD with the unbound structures of Ndt80 which vary from 0.83 Å (240 atoms-1MN4) to 2.67 Å (252 atoms - 1M6U-A). Although this is not altogether unexpected, due to the crystal packing constraints and changes induced by DNA binding, the extent to which the complex resists large structural changes is remarkable. What is unexpected is the trend that these mutant structures display. As the mutations and variations are made closer

to the middle of the MSE, where changes have the greatest effect on specificity and affinity, the RMSD actually decreases. One might expect that mutations that have the most detrimental effect on binding induce larger changes in the complex structure but the opposite effect is seen. The most significant changes in these structures as a whole tend to be on the non-prime strand where there are less protein contacts and rearrangements of the side chains that contact mutated base positions. In all cases the structural changes are localized around the sequence changes and do not propagate more than 2 base-pairs from the mutation.

**GC-rich region recognition**

The MSE can be crudely separated into two regions; a GC-rich region at one end of the consensus and a poly-A region at the other end. The GC-rich region is contacted in the major groove by Arg326, Arg111, and Arg177, residues that each recognize distinct YpG steps in the MSE. The TpG steps at positions 3/4 and 5/6 are well defined in the strict MSE consensus (Chu and Herskowitz, 1998; Hepworth et al., 1995; Ozsarac et al., 1997). The 5' pyrimidine of the CpG at -1/1 in our structure is only weakly conserved in the newly proposed MSE (5'-YGNCACAAAA-3') (Pierce et al., 2003). Five of our complexes test the consequences of disrupting these steps by either mutating the 5'-pyrimidine or the 3'-guanine. Mutation of the 3' guanine (seen in complexes vG1A/A9T, vG1C, and mC5T) is accommodated by rotation of the arginine away from the mutated base. In the case of the mutations at the position 1 guanine, the 5'-pyrimidine no

longer stacks on the arginine (Arg326) and shifts back towards the DNA duplex axis to stack on the 3'-guanine. This shift is in agreement with our proposal that the unstacked conformation seen in the wild type complex is energetically unfavorable in the absence of contacts with Ndt80. However, it is also likely that the new contact between Arg326 and the 5'-phosphate of the shifted pyrimidine also stabilizes this conformational change.

In contrast, mutation of any of the base pairs in the conserved YpG steps at either position 3/4 or 5/6 is not associated with a re-stacking of the DNA bases. One reason for this is that the region of the DNA from positions 3 to 6 has a large number of contacts to the protein in both the major and minor grooves as well as the DNA backbone. These extensive interactions likely anchor the DNA in its conformation with little regard for the changes made to the sequence. The energetic contributions of these contacts are high enough to compensate for the extra energy required to force the mutated dinucleotide step into the unstacked conformation, although this likely reduces the overall affinity of Ndt80 for these DNA substrates. It is interesting that the majority of these contacts are to only one strand of the DNA, the one that contains the YpG step, rather than its complementary strand.

The YpG step at positions 3 and 4 (gNCRCAAAA/T) contains an ambiguous YpG step originally thought to have no preference for a particular pyrimidine. However, recent work suggests that thymines are preferred over cytosine (Pierce et al., 2003). We have crystallized both the TpG (wild type) and CpG (vA4G) variations and the differences in these structures suggest an

explanation for this discrepancy. First, as described above, the standard base pairing geometry of the 5' CG base pair in the vA4G structure imposes a shift in the geometry of the backbone of the introduced guanine to the less stable BII conformation. Second, in the wild type structure the aliphatic portion of Arg177 forms a hydrophobic half pocket that cradles the 5-methyl group of the thymine (see Figure 4.2). We have shown that this methyl group is important for binding in the analogous TpG step in the 5$^{th}$ and 6$^{th}$ positions of the MSE with binding studies that show an approximately 2-fold reduction in binding affinity when this base is changed to a uracil (Lamoureux et al., 2002). When this thymine is mutated to a cytosine this base shifts slightly toward the major groove, possibly in an attempt to fill the hydrophobic "hole" left where the 5' methyl of the thymine would be. This shift further increases the unstacking of this step and likely increases the energy required to achieve this conformation, thus lowering the overall affinity.

**Poly-A tract recognition**

The second region of the MSE is the poly-A tract, which is recognized primarily through minor groove and backbone interactions. Because there are no major groove contacts to the bases past position 6 of the MSE, it was difficult to fully rationalize the selection of a poly-A tract in terms of direct readout. The minor groove interactions of Pro57 and Arg58 serve to exclude G-C or C-G base-pairs due to the steric clash that would occur with the 2-amino group of guanine, however the preference poly-A/poly-T over mixed A-T sequences was difficult to

explain based on the structure of the wild type complex alone. In the original paper of the wild type structure we predicted that preference for a poly-A tract might lie in the additional entropic cost of binding a flexible alternating A-T tract over a more rigid poly-A region that is pre-set in a conformation more suitable for Ndt80 binding. Several of the mutants crystallized here involve the poly-A tract and suggest an alternate structural mechanism that may account for this specificity.

At this point it is useful to consider the structure of poly-A tract DNA and some of its hallmarks. Poly-A DNA structures in the unbound state typically have a high propeller twist that enhances intra-strand base stacking as well as promoting inter-strand hydrogen bonding between adenine N6 and thymine O4 of adjacent base-pairs. Additionally, poly-A (as well as mixed A-T) regions tend to have an especially narrow minor groove of approximately 9.5 Å versus the 12 Å seen in B-DNA fibers (Nelson et al., 1987; Yoon et al., 1988). The poly-A region of the MSE DNA in the wild type complex does not exhibit high propeller twist nor is the minor groove narrowed, due to the insertion of the proline and arginine side-chains. In fact, at its widest point, near the 5' end, the minor groove of the poly-A tract is 14.1 Å. The widened minor groove brings the adenine N6 and thymine O4 of adjacent base-pairs closer together, such that high propeller twist is no longer necessary to establish hydrogen bonding between the adjacent base pairs. In fact, a high propeller twist in this region would cause the hydrogen bonding pair to clash.

To address the preference for poly-A over alternating A-T there are two mutants, mA7T and mA8T, which introduce an alternating A-T region and allow us to see the effect of this change. As is seen with many of these mutants, the DNA backbone has some small changes while by and large retaining the wild type structure. In both of these structures the largest structural changes are not observed in the mutated nucleotides but rather in the nucleotides 5' to the mutation on the otherwise poly-A strand. Note it is the poly-A strand that has fewer protein contacts and is expected to be more flexible than the poly-T strand. The adenosine 5' to the mutated base shifts 0.7 Å toward the minor groove in the mA7T structure and 0.4 Å in mA8T structure. This shift helps to reduce steric repulsions between the two N6 atoms of adenines of adjacent base-pairs at the introduced 5'-ApT-3' step. Despite this rearrangement, the N6 atoms are still close enough to repel one another. The N6 - N6 distances are 3.0 and 3.1 Å in the mA7T and mA8T structures respectively. This steric clash likely accounts for the lower binding affinity of these mutant MSEs (Figure 4.8 and 4.9). This effect is greatest at the 5' end of the poly-A tract (position 6-7) where the minor groove is the widest and diminishes toward the 3' end so that at position 9, where the minor groove width is similar to unbound poly-A DNA (9.5 Å at position 9), there is little or no preference for A-T over T-A base-pairs.

In addition to the N6 clash of adjacent adenines in mA8T, the mutated thymine also introduces a potential clash between its 5-methyl group and its 5' phosphate. This clash is alleviated by the backbone adopting a more "BI-like" conformation that shifts the phosphate group away from the 5-methyl, which has

an additional effect of moving the 5' adenine in the right direction to minimize the N6-N6 clash.

**Implications for modeling meiotic transcriptional activation in *S. cerevisiae***

The structures presented here provide a detailed view of how Ndt80 binds a number of different DNA targets with subtle changes that modulate binding affinities within an order of magnitude. Classic experiments on the λ phage repressor/cro system have provided the best-known example of how differences in affinity of transcription factors for different DNA targets can drive a developmental program (in this case the switch from lysogenic to lytic growth) in response to changing levels of transcription factor concentrations (Ptashne, 1986). It is tempting to speculate that a similar mechanism may regulate the precise timing of gene expression during progression through meiosis, dependent on the relative binding affinities of Ndt80 and Sum1 for key regulatory MSE elements (Pierce et al., 2003; Xie et al., 1999). For example, the activation of certain genes very soon after the recombination checkpoint might be explained by a relatively high affinity of Ndt80 and/or a relatively low affinity for Sum1 in key regulatory MSEs. Conversely, activation of other genes could be delayed until later in the developmental program, when Ndt80 protein levels are higher and Sum1 is lower, by the utilization of MSE elements which have a correspondingly lower affinity for Ndt80 and/or higher affinity for Sum1. This differential timing of middle genes depends on whether these genes have a Sum1 binding site, an Ndt80 binding site, or both. In addition, the relative affinities of these binding sites

can further influence the timing and result in activation of middle genes over the entire time course of middle sporulation. In fact, the four waves of gene expression originally used to describe sporulation were expanded to seven in a DNA microarray study that focused on sporulation (Chu et al., 1998). This work indicates that, as the transcription program is investigated on a finer scale, more temporal patterns will emerge.

Ndt80 and Sum1 have overlapping although distinct MSE binding site requirements and that the binding of one of these proteins to its MSE is mutually exclusive (Pierce et al., 2003). These properties combined with the fact that Ndt80 auto-induces its own expression in a positive feedback loop have been used in computer models to generate a network with an extremely sharp expression profile (Wang et al., 2005). It has been speculated that both auto-feedback and activator/repressor competition may be general features of processes that require sharp temporal and/or spatial gene expression, such as sporulation in yeast or developmental/differentiation pathways of higher eukaryotes (Wang et al., 2005). If this is the case, then the Ndt80/Sum1/MSE system could provide an ideal model to study this type of regulatory network.

The sharp expression profile of Sum1/Ndt80 dual regulated middle genes, in addition to the potential differential timing of middle genes, forms an interesting and complicated expression system. *In vivo* this system is further complicated by controls at the level of protein synthesis and/or degradation (Hepworth et al., 1998; Lindgren et al., 2000). Furthermore there are post-translational modifications that change the activity of these transcription factors, possibly in

response to recombination checkpoints(Tung et al., 2000). As well there may be other transcription factors that enhance or antagonize the activity of Ndt80 or Sum1. If we hope to understand this system as a whole *in vivo*, a first step is to understand the basic principles of the DNA - protein recognition.

**Conclusions**

We have demonstrated the importance of protein contacts to the DNA backbone in molding the conformation of the DNA, even when the MSE is not conserved. This highlights the contribution of indirect readout on MSE recognition. The critical recognition of YpG steps in the 5' part of the MSE is in part indirect readout of the intrinsic flexibility of the pyrimidine-purine step, and direct readout of the base pair surfaces exposed in the major groove through hydrogen bonding, cation-$\pi$ and van der Waals interactions. Indirect readout seems to be implicated in poly-A tract recognition as well. It appears that Ndt80 does not specifically recognize the conformation of the poly-A tract in its unbound state; rather, this region of the MSE is recognized by its ability to adopt a conformation induced by Ndt80 binding while simultaneously avoiding steric clashes within the DNA and with the protein. The adjustments in the DNA are often accommodated by subtle shifts between BI and BII backbone conformations, allowing flexure of the double helix to adapt to its protein partner.

## REFERENCES


Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Chu, S., and Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. Mol Cell *1*, 685-696.

Collaborative Computational Project, N. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr *50*, 760-763.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System (San Carlos, CA, USA: DeLano Scientific).

Fingerman, I. M., Sutphen, K., Montano, S. P., Georgiadis, M. M., and Vershon, A. K. (2004). Characterization of critical interactions between Ndt80 and MSE DNA defining a novel family of Ig-fold transcription factors. Nucleic Acids Res *32*, 2947-2956.

Hepworth, S. R., Ebisuzaki, L. K., and Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in Saccharomyces cerevisiae. Mol Cell Biol *15*, 3934-3944.

Hepworth, S. R., Friesen, H., and Segall, J. (1998). NDT80 and the meiotic recombination checkpoint regulate expression of middle sporulation-specific genes in Saccharomyces cerevisiae. Mol Cell Biol *18*, 5750-5761.

Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A *47 ( Pt 2)*, 110-119.

Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C., and Glover, J. N. M. (2002). Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. Embo J *21*, 5721-5732.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst, 283-291.

Lindgren, A., Bungard, D., Pierce, M., Xie, J., Vershon, A., and Winter, E. (2000). The pachytene checkpoint in Saccharomyces cerevisiae requires the Sum1 transcriptional repressor. Embo J *19*, 6489-6497.

Lu, X. J., Shakked, Z., and Olson, W. K. (2000). A-form conformational motifs in ligand-bound DNA structures. J Mol Biol *300*, 819-840.

Montano, S. P., Cote, M. L., Fingerman, I., Pierce, M., Vershon, A. K., and Georgiadis, M. M. (2002). Crystal structure of the DNA-binding domain from Ndt80, a transcriptional activator required for meiosis in yeast. Proc Natl Acad Sci U S A *99*, 14041-14046.

Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., and Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A *95*, 11163-11168.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode, In Methods in Enzymology, C. W. Carter Jr., and R. M. Sweet, eds. (New York: Academic Press), pp. 307-326.

Ozsarac, N., Straffon, M. J., Dalton, H. E., and Dawes, I. W. (1997). Regulation of gene expression during meiosis in Saccharomyces cerevisiae: SPR3 is controlled by both ABFI and a new sporulation control element. Mol Cell Biol *17*, 1152-1159.

Pierce, M., Benjamin, K. R., Montano, S. P., Georgiadis, M. M., Winter, E., and Vershon, A. K. (2003). Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol *23*, 4814-4825.

Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L., and Dickerson, R. E. (1987). Helix geometry, hydration, and G.A mismatch in a B-DNA decamer. Science *238*, 498-504.

Ptashne, M. (1986). A genetic switch : gene control and phage [lamda] (Palo Alto, CA: Blackwell Scientific Publications & Cell Press).

Tung, K. S., Hong, E. J., and Roeder, G. S. (2000). The pachytene checkpoint prevents accumulation and phosphorylation of the meiosis-specific transcription factor Ndt80. Proc Natl Acad Sci U S A *97*, 12187-12192.

Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D., and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci U S A *102*, 1998-2003.

Xie, J., Pierce, M., Gailus-Durner, V., Wagner, M., Winter, E., and Vershon, A. K. (1999). Sum1 and Hst1 repress middle sporulation-specific gene expression during mitosis in Saccharomyces cerevisiae. Embo J *18*, 6448-6454.

Yoon, C., Prive, G. G., Goodsell, D. S., and Dickerson, R. E. (1988). Structure of an alternating-B DNA helix and its relationship to A-tract DNA. Proc Natl Acad Sci U S A *85*, 6332-6336.

# Chapter 5:

# Discussion and Conclusions

**Ig-fold transcription factors**

The family of Ig-fold transcription factors includes such notable names as; p53 (Cho et al., 1994), STAT (Becker et al., 1998), NF-κB (Cramer et al., 1997), NFAT (Chen et al., 1998), Runt (Bravo et al., 2001; Tahirov et al., 2001) and the Brachyhury T-box (Muller and Herrmann, 1997) sub-families. All of these proteins are found in metazoans and they all play a somewhat similar role in either developmental pathways or the immune response. In fact many of these proteins were discovered because of the aberrant phenotype displayed when they are mutated, typically tumor formation or developmental defects. These severe phenotypes hint at the critical nature of these proteins to cellular development and survival. The structure of Ndt80 is the first example of this fold found in a non-metazoan organism and maintains the theme of functioning in developmental pathways. In the case of yeast the developmental pathway is meiosis, which Ndt80 plays a large part in coordinating, particularly during the middle phase.

**Structural similarities**

An interesting note on Ig-fold transcription factors is that they all appear to be members of the s-type Ig-fold family (Figure 5.1) (Rudolph and Gergen, 2001) . While the 4 subfamilies of Ig folds (s, h, c, v - types) are all very similar, why the transcription factors only fall into one category is uncertain. In fact, the only difference between these sub families lies between the c and e strands. These four sub-families are distinguished only by the

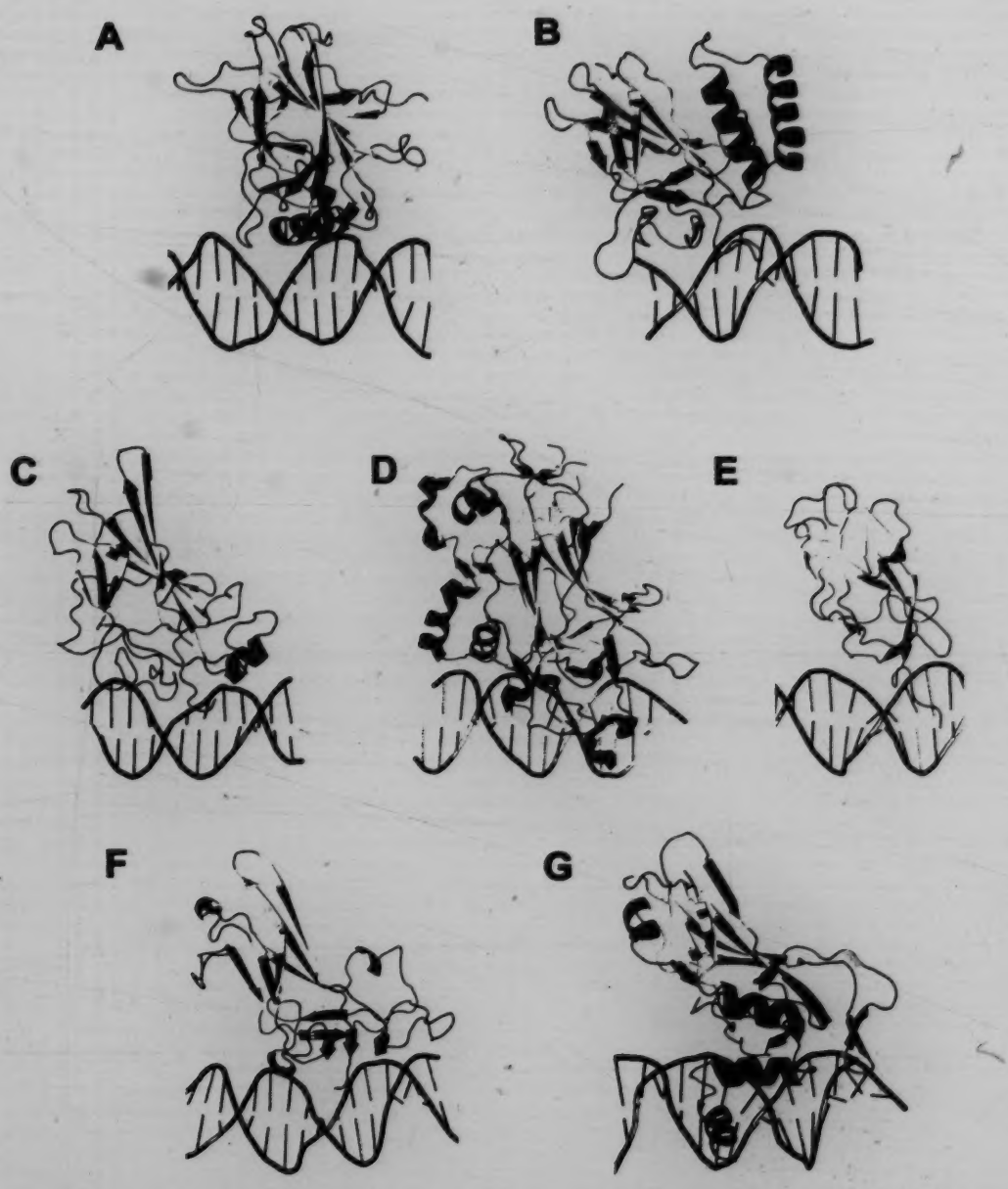


**Figure 5.1. Comparison of Ig-fold transcription factors.** The core Ig fold is colored yellow, the AB loop red, EF loop blue and C-terminal region green. Each family is represented with the PDB used to generate the image in brackets **(A)** p53 (1TSR) **(B)** p52/Rel family (1A3Q) **(C)** Nfat (1A66) **(D)** Ndt80 (1MNN) **(E)** Runt domain (1HJC) **(F)** Stat family (1BG1) **(G)** T-box domain (1XBR)

existence and position of a c', c" and/or d strand (Bork et al., 1994). Even more striking is that all of these transcription factors use the same elements to recognize their DNA targets, namely the AB loop, the EF loop, and the region C-terminal to the Ig fold. In addition, the orientation of these 3 recognition loops are always the same relative to each other, with the C-terminal region sandwiched between the AB and EF loops. This suggests that all these transcription factors may be evolutionarily related, despite their exceptionally low degree of sequence similarity.

**The possibility of evolutionary links**

The poor degree of sequence conservation does not seem to be a restriction to an evolutionary relationship. In general the strands that make the Ig-core have very different primary sequences and there is little selective pressure to prevent mutation as long as the hydrophobic core remains intact. A specific example is p53. Standard searches of genomic sequences did not find a p53 homolog in *C. elegans*. However, when additional algorithms were used to consider p53 signature sequences and residues most commonly mutated in human cancers, a sequence proposed to be the p53 homolog was discovered (Derry et al., 2001). This homolog, Cep-1, was shown to be functionally similar to p53 (Derry et al., 2001) and later its structure was determined to have a remarkable degree of similarity to the human p53 (RMSD of 4.0 Å over all Cα atoms and less than 3.0 Å for most Ig core

residues) although the primary sequence identity was only 15% (Huyen et al., 2004).

The above example of mammalian p53 and Cep-1 sets a precedent which can be used to suggest an evolutionary link between Ndt80 and p53 despite the lack of sequence conservation. Cep-1 not only mediates the cell response to DNA damage like vertebrate p53, but is also essential for proper segregation of chromosomes in meiosis I (Derry et al., 2001). Like Ndt80, Cep-1 is involved in the progression through meiosis. Although mammalian p53 appears to be dispensable for meiosis (Gersten and Kemp, 1997), there are high levels of p53 expression in tetraploid primary spermatocytes during pachytene indicating there may be a non essential role for p53 during meiosis (Schwartz et al., 1993; Sjoblom and Lahdetie, 1996). This link lends credibility to the idea that checkpoints regulating cell cycle progression in response to DNA damage - either from meiotic recombination or exogenous causes - may share an evolutionary relationship.

**Transcriptional regulation of meiosis**

Most of the genes that are coordinately regulated during the middle phase of sporulation contain an MSE sequence within their promoters (Chu et al., 1998). Ndt80 and Sum1 have overlapping though distinct MSE binding preferences (Pierce et al., 2003), thus the variation in the composition of these MSE sequences and possibly of the flanking sequences can have a large effect on its regulatory activity during meiosis. One can imagine the

relative affinities of the MSE for either Sum1 or Ndt80 can result in the different magnitudes and timing of middle gene expression seen in gene chip experiments (Chu et al., 1998). MSE sites can be broadly classified as vegetative repressor elements bound primarily by Sum1, meiotic activator elements bound by Ndt80 or switch elements which have high affinities for both proteins (Pierce et al., 2003). It is these switch elements that are of particular interest.

Ndt80 and Sum1 were used to construct a computer model of the transcriptional network architecture of these two proteins in context with the Ndt80 MSE site (a switch element MSE). As mentioned in the introduction, this type of network results in an exceptionally sharp and temporally distinct expression profile. In the model, binding site competition between activator and repressor and auto-feedback are required characteristics for this behavior, see Figure 5.2. This model is intriguing as it suggests that developmental pathways that require sharp expression profiles (both temporally and spatially) may share these characteristics. Although this model neglects to consider post transcriptional modifications, if valid, it may provide the basis for many transcriptional networks. Using the Ndt80-Sum1-MSE system and a reporter construct this model could be validated by substituting single positions along the MSE in the reporter. If the changes in the timing and magnitude of expression correlate to those predicted it would provide a solid basis for this model.

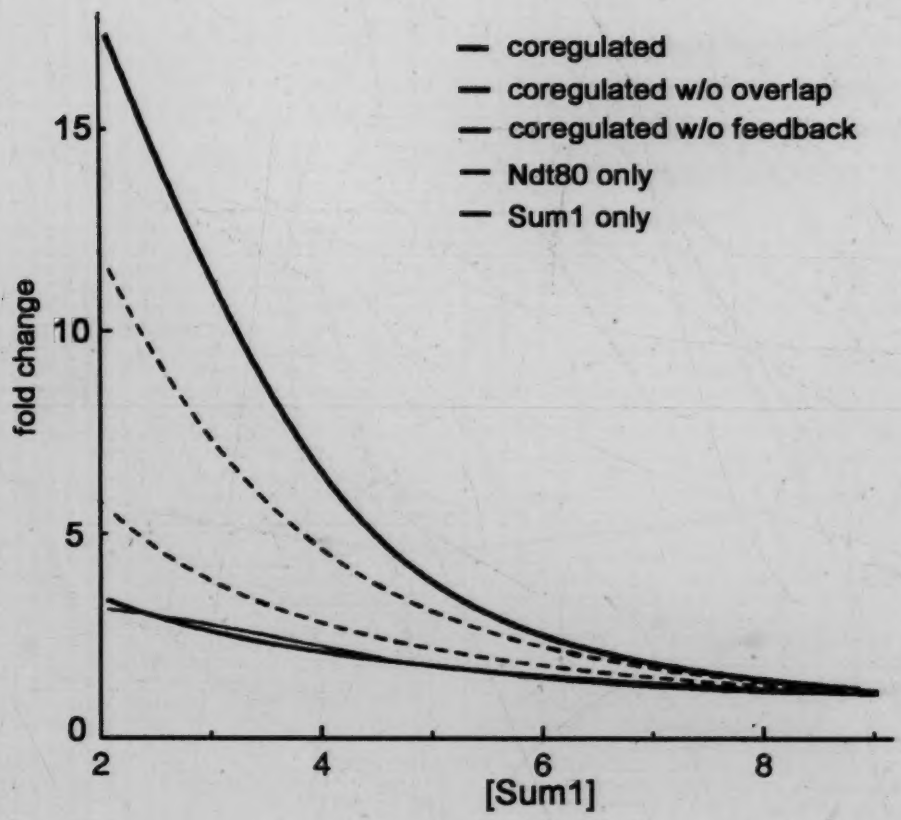


**Figure 5.2 Transcriptional activation versus active Sum1 concentrations in a model system.** The red, orange and black solid lines represent Sum1 only, Ndt80 only, and dually regulated genes respectively. The dashed magenta line represents the hypothetical dually regulated gene where there is no consensus site overlap. The dashed blue line represents dually regulated gene in a system without positive auto feedback. This graph is reconstructed from the model presented by (Wang et al., 2005).

Of course this simple binary system is not sufficient to explain the real expression profiles observed. For instance, Ndt80 expression is complicated by other factors such as Ume6-Sin3-Rpd3 repressor complex. In *sum1* mutants where repression is expected to be released, NDT80 is not transcribed, presumably due to the to URS1 sequences that recruit the Ume6 repression complex. Furthermore, transcription factor binding sites adjacent to the MSE, such as for Abf1, have been implicated in co-operatively increasing transcription of a reporter during meiosis (Hepworth et al., 1995) This co-operativity does not appear to be at the level of DNA binding since electro-mobility shift assays of complexes containing DNA with both consensus sites, Ndt80 and Abf1 is not co-operative (data not shown). Clearly the *in vivo* system is much more complicated than the model referred to but nonetheless the model may provide a useful starting point upon which to build.

**DNA-protein recognition principles**

Chapters 3 and 4 of this thesis primarily deal with DNA - protein recognition of Ndt80 - MSE. A novel mode of recognition was described for YpG steps in the DNA and evidence was provided for its role in consensus recognition of numerous other transcription factors. Although this mode of recognition appears to be utilized by almost all major families of transcription factors it is not universal. That is not all instances of YpG steps recognized by an arginine residue will unstack its bases to facilitate stacking with the guanidinium group. The reason is that this type of recognition is a concerted

effort of indirect and direct contacts. The energy gained through the cation-π interaction is not sufficient to distort dinucleotide step. However, when combined with the backbone contacts which stabilize the nonsymmetrical unstacking of this step this cation-π interaction becomes important for specificity.

**On "recognition codes"**

Similar to the Zif268 study, no "recognition code" can be assigned to Ndt80 - MSE interactions. In fact, much of the "recognition" of the MSE is restricting the possible base pairs that can occupy a particular position through either steric or conformational constraints rather that direct recognition of elements that are present. For instance the poly-A tract in the MSE is only contacted through the minor groove and backbone interactions. These minor groove interactions exclude G-C or C-G base pairs from occupying these positions through a steric clash of the 2 - amino group of the guanine. In addition T-A base pairs are excluded from these positions by the conformation of the DNA helix in this vicinity. The minor groove is widened such that replacing an A-T base pair with a T-A would cause a clash of the N6 atoms of adjacent adenosines on opposite strands (as in Fig 4.7). In the Ndt80-MSE complex there are only 3 canonical direct contacts to a consensus site although there are base preferences in at least 8 positions. This trend is common among DNA-protein complexes. In cases where the consensus DNA target is well defined and the structure has been solved there

is commonly a discrepancy between *in vivo* specificity versus the specificity that is inferred by structural analysis. In some cases this deficiency may be due to the action of other proteins but many of these systems have *in vitro* binding experiments that determine the proteins consensus preference in isolation. Therefore indirect readout must play a significant role. Because the mechanisms of indirect readout are typically much more subtle and often context specific within a system, universal rules for DNA-protein interactions seems like an unreasonable goal.

**Technical issues of DNA modeling**

One development that may help in the future refinement of DNA and DNA-protein complexes is to revisit the geometric constraints and weights of different DNA conformations utilized by refinement software, particularly the BI and BII forms of DNA. On multiple occasions regions of the DNA that were modeled in the BII conformation reverted to a BI conformation with a single round of refinement. These regions had clear, high resolution electron density in which the refined BI conformation clearly came out of density. This indicates that the refinement programs tend to outweigh the BI versus BII conformation to the point that experimental data (i.e. the X-ray diffraction intensities) essentially becomes outweighed. This is of less concern with high resolution structures, where the correct conformation can be manually built, but with lower resolution structures, where the conformation is ambiguous, this may skew the vast majority of DNA into BI conformation.

**Conclusions**

Despite the lack of canonical direct readout we have reasonable explanations for the specificity of every position that has been shown to be selected for. Although this sounds like a mundane claim there are remarkably few protein-DNA complexes that can make this claim and even fewer with supporting structural work. Our structural explanations for specificity and affinity could be strengthened by the addition of energetic values, such as those derived from ITC experiments. Such experiments could supply values of the relative enthalpic and entropic costs of the mutants described in Chapter 4 and potentially point out new features to account for the affinity and specificity differences between these mutants.

Ndt80 - MSE - Sum1 system is well on its way to becoming a model system for understanding gene regulation in developmental pathways as well as general DNA - protein interactions. The plethora of data on this system in the areas of: promoter analysis, gene expression via microarray analysis, *in vitro* binding assays, mutagenesis of both Ndt80 and the MSE, and finally structural analysis all help to establish important theories in this field. The use of simple physical chemistry models has indicated that the properties of competition and autofeedback inherent in the system may be requirements of these sharp transition transcriptional networks (Wang et al., 2005). A full understanding of the DNA - protein interactions is a critical building block required to comprehend these transcriptional regulatory pathways.

## REFERENCES


Becker, S., Groner, B., and Muller, C. W. (1998). Three-dimensional structure of the Stat3beta homodimer bound to DNA. Nature *394*, 145-151.

Bork, P., Holm, L., and Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol *242*, 309-320.

Bravo, J., Li, Z., Speck, N. A., and Warren, A. J. (2001). The leukemia-associated AML1 (Runx1)--CBF beta complex functions as a DNA-induced molecular clamp. Nat Struct Biol *8*, 371-378.

Chen, L., Glover, J. N. M., Hogan, P. G., Rao, A., and Harrison, S. C. (1998). Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. Nature *392*, 42-48.

Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346-355.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Cramer, P., Larson, C. J., Verdine, G. L., and Muller, C. W. (1997). Structure of the human NF-kappaB p52 homodimer-DNA complex at 2.1 A resolution. Embo J *16*, 7078-7090.

Derry, W. B., Putzke, A. P., and Rothman, J. H. (2001). Caenorhabditis elegans p53: role in apoptosis, meiosis, and stress resistance. Science *294*, 591-595.

Gersten, K. M., and Kemp, C. J. (1997). Normal meiotic recombination in p53-deficient mice. Nat Genet *17*, 378-379.

Hepworth, S. R., Ebisuzaki, L. K., and Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in Saccharomyces cerevisiae. Mol Cell Biol *15*, 3934-3944.

Huyen, Y., Jeffrey, P. D., Derry, W. B., Rothman, J. H., Pavletich, N. P., Stavridi, E. S., and Halazonetis, T. D. (2004). Structural differences in the DNA binding domains of human p53 and its C. elegans ortholog Cep-1. Structure (Camb) *12*, 1237-1243.

Muller, C. W., and Herrmann, B. G. (1997). Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. Nature *389*, 884-888.

Pierce, M., Benjamin, K. R., Montano, S. P., Georgiadis, M. M., Winter, E., and Vershon, A. K. (2003). Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol 23, 4814-4825.

Rudolph, M. J., and Gergen, J. P. (2001). DNA-binding by Ig-fold proteins. Nat Struct Biol *8*, 384-386.

Schwartz, D., Goldfinger, N., and Rotter, V. (1993). Expression of p53 protein in spermatogenesis is confined to the tetraploid pachytene primary spermatocytes. Oncogene *8*, 1487-1494.

Sjoblom, T., and Lahdetie, J. (1996). Expression of p53 in normal and gamma-irradiated rat testis suggests a role for p53 in meiotic recombination and repair. Oncogene *12*, 2499-2505.

Tahirov, T. H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., *et al.* (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell *104*, 755-767.

Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D., and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci U S A *102*, 1998-2003.

## Appendix A
## List of PDB used in Chapter 3 as representative of the DNA-protein complex database

pdb10mh.ent
pdb1a02.ent
pdb1a0a.ent
pdb1a1f.ent
pdb1a1g.ent
pdb1a1h.ent
pdb1a1i.ent
pdb1a1j.ent
pdb1a1k.ent
pdb1a1l.ent
pdb1a1v.ent
pdb1a31.ent
pdb1a35.ent
pdb1a36.ent
pdb1a3q.ent
pdb1a6y.ent
pdb1a73.ent
pdb1a74.ent
pdb1aay.ent
pdb1ais.ent
pdb1akh.ent
pdb1am9.ent
pdb1an2.ent
pdb1an4.ent
pdb1aoi.ent
pdb1apl.ent
pdb1au7.ent
pdb1awc.ent
pdb1az0.ent
pdb1azp.ent
pdb1azq.ent
pdb1b01.ent
pdb1b3t.ent
pdb1b72.ent
pdb1b8i.ent
pdb1b94.ent
pdb1b95.ent
pdb1b96.ent
pdb1b97.ent
pdb1bc7.ent
pdb1bc8.ent
pdb1bdh.ent
pdb1bdi.ent
pdb1bdt.ent
pdb1bdv.ent
pdb1bf4.ent
pdb1bf5.ent
pdb1bg1.ent
pdb1bgb.ent
pdb1bhm.ent
pdb1bl0.ent
pdb1bnk.ent
pdb1bnz.ent
pdb1bp7.ent
pdb1bpx.ent
pdb1bpy.ent
pdb1bpz.ent
pdb1brn.ent
pdb1bss.ent
pdb1bsu.ent
pdb1bua.ent
pdb1bvo.ent
pdb1c8c.ent
pdb1c9b.ent
pdb1ca5.ent
pdb1ca6.ent
pdb1cbv.ent
pdb1cdw.ent
pdb1cez.ent
pdb1cf7.ent
pdb1cgp.ent
pdb1cit.ent
pdb1ckq.ent
pdb1ckt.ent
pdb1cl8.ent
pdb1clq.ent
pdb1cma.ent
pdb1crx.ent
pdb1cw0.ent
pdb1cyq.ent
pdb1cz0.ent
pdb1d02.ent
pdb1d0e.ent
pdb1d1u.ent
pdb1d2i.ent
pdb1d3u.ent
pdb1d5y.ent
pdb1d66.ent
pdb1d8y.ent
pdb1dc1.ent
pdb1dct.ent
pdb1ddn.ent
pdb1de8.ent
pdb1de9.ent
pdb1dew.ent
pdb1dfm.ent
pdb1dgc.ent
pdb1dh3.ent
pdb1diz.ent
pdb1dmu.ent
pdb1dnk.ent
pdb1dp7.ent
pdb1drg.ent
pdb1dsz.ent
pdb1du0.ent
pdb1dux.ent
pdb1e3m.ent
pdb1e3o.ent
pdb1ea4.ent
pdb1ebm.ent
pdb1ecr.ent
pdb1efa.ent
pdb1egw.ent
pdb1ehl.ent
pdb1ej9.ent
pdb1emh.ent
pdb1emj.ent
pdb1eo3.ent
pdb1eo4.ent
pdb1eon.ent
pdb1eoo.ent
pdb1eop.ent
pdb1eqz.ent
pdb1eri.ent
pdb1esg.ent
pdb1ewn.ent
pdb1ewq.ent
pdb1exj.ent
pdb1eyg.ent
pdb1eyu.ent
pdb1f0o.ent
pdb1f0v.ent
pdb1f2i.ent
pdb1f44.ent
pdb1f4k.ent
pdb1f4r.ent
pdb1f5t.ent
pdb1f66.ent
pdb1f6o.ent
pdb1fiu.ent
pdb1fjl.ent
pdb1fjx.ent
pdb1flo.ent
pdb1fn7.ent
pdb1fok.ent
pdb1fw6.ent
pdb1fyk.ent
pdb1fyl.ent
pdb1fym.ent
pdb1fzp.ent
pdb1g2d.ent
pdb1g2f.ent
pdb1g38.ent
pdb1g3x.ent
pdb1g9y.ent
pdb1g9z.ent
pdb1ga5.ent
pdb1gd2.ent
pdb1gdt.ent
pdb1gji.ent
pdb1glu.ent
pdb1gt0.ent
pdb1gu4.ent
pdb1gu5.ent
pdb1gxp.ent
pdb1h0m.ent
pdb1h6f.ent
pdb1h88.ent

pdb1h89.ent
pdb1h8a.ent
pdb1h9d.ent
pdb1hao.ent
pdb1hap.ent
pdb1hcq.ent
pdb1hcr.ent
pdb1hdd.ent
pdb1hf0.ent
pdb1hht.ent
pdb1hi0.ent
pdb1hjb.ent
pdb1hjc.ent
pdb1hlo.ent
pdb1hlv.ent
pdb1hlz.ent
pdb1hu0.ent
pdb1huo.ent
pdb1hut.ent
pdb1huz.ent
pdb1hwt.ent
pdb1i3j.ent
pdb1i6j.ent
pdb1i7d.ent
pdb1i8m.ent
pdb1iaw.ent
pdb1ic8.ent
pdb1if1.ent
pdb1ig7.ent
pdb1ig9.ent
pdb1ign.ent
pdb1ihf.ent
pdb1ijw.ent
pdb1imh.ent
pdb1io4.ent
pdb1ipp.ent
pdb1iu3.ent
pdb1ixy.ent
pdb1j1v.ent
pdb1j59.ent
pdb1j75.ent
pdb1jb7.ent
pdb1je8.ent
pdb1jey.ent
pdb1jfi.ent
pdb1jfs.ent
pdb1jft.ent
pdb1jgg.ent
pdb1jh9.ent
pdb1jj4.ent
pdb1jj6.ent
pdb1jj8.ent
pdb1jk1.ent
pdb1jk2.ent
pdb1jko.ent
pdb1jkp.ent
pdb1jkq.ent
pdb1jkr.ent
pdb1jmc.ent
pdb1jnm.ent
pdb1jt0.ent
pdb1jx4.ent
pdb1jxl.ent
pdb1k3w.ent
pdb1k3x.ent
pdb1k4t.ent
pdb1k61.ent
pdb1k78.ent
pdb1k79.ent
pdb1k7a.ent
pdb1k82.ent
pdb1k8g.ent
pdb1kb2.ent
pdb1kb4.ent
pdb1kb6.ent
pdb1kbu.ent
pdb1kc6.ent
pdb1kdh.ent
pdb1keg.ent
pdb1kfs.ent
pdb1kfv.ent
pdb1kix.ent
pdb1krp.ent
pdb1ksp.ent
pdb1ku7.ent
pdb1kx3.ent
pdb1kx4.ent
pdb1kx5.ent
pdb1l1t.ent
pdb1l1z.ent
pdb1l2b.ent
pdb1l2c.ent
pdb1l2d.ent
pdb1l3l.ent
pdb1l3s.ent
pdb1l3t.ent
pdb1l3u.ent
pdb1l3v.ent
pdb1l5u.ent
pdb1lat.ent
pdb1lau.ent
pdb1le5.ent
pdb1le8.ent
pdb1le9.ent
pdb1lei.ent
pdb1lli.ent
pdb1lmb.ent
pdb1lq1.ent
pdb1lrr.ent
pdb1lv5.ent
pdb1lwv.ent
pdb1lww.ent
pdb1lwy.ent
pdb1m07.ent
pdb1m0e.ent
pdb1m18.ent
pdb1m19.ent
pdb1m1a.ent
pdb1m5r.ent
pdb1m5x.ent
pdb1m6x.ent
pdb1ma7.ent
pdb1mdm.ent
pdb1mdy.ent
pdb1mey.ent
pdb1mhd.ent
pdb1mht.ent
pdb1mj2.ent
pdb1mjm.ent
pdb1mjo.ent
pdb1mjq.ent
pdb1mm8.ent
pdb1mnm.ent
pdb1mnn.ent
pdb1mnv.ent
pdb1mow.ent
pdb1mq3.ent
pdb1mtl.ent
pdb1muh.ent
pdb1mur.ent
pdb1mus.ent
pdb1mwi.ent
pdb1mwj.ent
pdb1n39.ent
pdb1n3a.ent
pdb1n3c.ent
pdb1n3e.ent
pdb1n3f.ent
pdb1n4l.ent
pdb1n6q.ent
pdb1nfk.ent
pdb1ng9.ent
pdb1ngm.ent
pdb1nkp.ent
pdb1nlw.ent
pdb1nnj.ent
pdb1noy.ent
pdb1nwq.ent
pdb1o3q.ent
pdb1o3r.ent
pdb1o3s.ent
pdb1o3t.ent
pdb1oct.ent
pdb1odg.ent
pdb1odh.ent
pdb1oe4.ent
pdb1oe5.ent
pdb1oe6.ent
pdb1oh5.ent
pdb1oh6.ent
pdb1oh7.ent
pdb1oh8.ent
pdb1orn.ent
pdb1orp.ent
pdb1otc.ent
pdb1oup.ent
pdb1ouz.ent
pdb1owf.ent
pdb1owg.ent
pdb1p47.ent
pdb1p4e.ent
pdb1p51.ent
pdb1p59.ent
pdb1p71.ent

pdb1p78.ent
pdb1p7d.ent
pdb1pa6.ent
pdb1par.ent
pdb1pdn.ent
pdb1per.ent
pdb1ph1.ent
pdb1ph3.ent
pdb1ph4.ent
pdb1ph5.ent
pdb1ph6.ent
pdb1ph7.ent
pdb1ph8.ent
pdb1ph9.ent
pdb1phj.ent
pdb1pnr.ent
pdb1pue.ent
pdb1puf.ent
pdb1pv4.ent
pdb1pvi.ent
pdb1qai.ent
pdb1qaj.ent
pdb1qbj.ent
pdb1qn3.ent
pdb1qn4.ent
pdb1qn5.ent
pdb1qn6.ent
pdb1qn7.ent
pdb1qn8.ent
pdb1qn9.ent
pdb1qna.ent
pdb1qnb.ent
pdb1qnc.ent
pdb1qne.ent
pdb1qp0.ent
pdb1qp4.ent
pdb1qp7.ent
pdb1qp9.ent
pdb1qpi.ent
pdb1qps.ent
pdb1qpz.ent
pdb1qqa.ent
pdb1qqb.ent
pdb1qrh.ent
pdb1qri.ent
pdb1qrv.ent
pdb1qsl.ent
pdb1qss.ent
pdb1qsy.ent
pdb1qtm.ent
pdb1qum.ent
pdb1ram.ent
pdb1rbj.ent
pdb1rcn.ent
pdb1rep.ent
pdb1rnb.ent
pdb1rpe.ent
pdb1rta.ent
pdb1run.ent
pdb1ruo.ent
pdb1rv5.ent
pdb1rva.ent
pdb1rvb.ent
pdb1rvc.ent
pdb1skn.ent
pdb1ssp.ent
pdb1svc.ent
pdb1t7p.ent
pdb1tau.ent
pdb1tc3.ent
pdb1tgh.ent
pdb1tro.ent
pdb1trr.ent
pdb1tsr.ent
pdb1tup.ent
pdb1uaa.ent
pdb1ubd.ent
pdb1vas.ent
pdb1vkx.ent
pdb1vol.ent
pdb1vpw.ent
pdb1wet.ent
pdb1xbr.ent
pdb1yrn.ent
pdb1ysa.ent
pdb1ytb.ent
pdb1ytf.ent
pdb1zaa.ent
pdb1zay.ent
pdb1zme.ent
pdb1zqa.ent
pdb1zqf.ent
pdb1zqi.ent
pdb1zqn.ent
pdb1zqp.ent
pdb2bam.ent
pdb2bdp.ent
pdb2bop.ent
pdb2bpa.ent
pdb2bpf.ent
pdb2cgp.ent
pdb2crx.ent
pdb2dgc.ent
pdb2dnj.ent
pdb2drp.ent
pdb2glj.ent
pdb2hap.ent
pdb2hdd.ent
pdb2hmi.ent
pdb2irf.ent
pdb2kfn.ent
pdb2kfz.ent
pdb2ktq.ent
pdb2kzm.ent
pdb2kzz.ent
pdb2nll.ent
pdb2or1.ent
pdb2pjr.ent
pdb2pua.ent
pdb2pub.ent
pdb2puc.ent
pdb2pud.ent
pdb2pue.ent
pdb2puf.ent
pdb2pug.ent
pdb2pvi.ent
pdb2ram.ent
pdb2rve.ent
pdb2ssp.ent
pdb2up1.ent
pdb3bam.ent
pdb3bdp.ent
pdb3cro.ent
pdb3crx.ent
pdb3hdd.ent
pdb3hts.ent
pdb3ktq.ent
pdb3mht.ent
pdb3orc.ent
pdb3pvi.ent
pdb4bdp.ent
pdb4crx.ent
pdb4dpv.ent
pdb4ktq.ent
pdb4mht.ent
pdb4rve.ent
pdb4skn.ent
pdb5crx.ent
pdb5mht.ent
pdb6cro.ent
pdb6mht.ent
pdb6pax.ent
pdb7ice.ent
pdb7icg.ent
pdb7ich.ent
pdb7ici.ent
pdb7ick.ent
pdb7icm.ent
pdb7icn.ent
pdb7icp.ent
pdb7icq.ent
pdb7icr.ent
pdb7ics.ent
pdb7ict.ent
pdb7icv.ent
pdb7mht.ent
pdb8ica.ent
pdb8icc.ent
pdb8icf.ent
pdb8ici.ent
pdb8ick.ent
pdb8icm.ent
pdb8icn.ent
pdb8ico.ent
pdb8icp.ent
pdb8icq.ent
pdb8icr.ent
pdb8ics.ent
pdb8icu.ent
pdb8icx.ent
pdb8mht.ent
pdb9ant.ent
pdb9ica.ent
pdb9icf.ent

pdb9icg.ent
pdb9ich.ent
pdb9ick.ent
pdb9icl.ent
pdb9icm.ent
pdb9icn.ent
pdb9ico.ent
pdb9icq.ent
pdb9icr.ent
pdb9ics.ent
pdb9ict.ent
pdb9icu.ent
pdb9icv.ent
pdb9icw.ent
pdb9icx.ent
pdb9icy.ent
pdb9mht.ent

**Appendix B**
**Perl Script used to parse the database (Appendix A)**

```
#!/usr/bin/perl
use Math::Trig;
#################################################################
##      Program: ArgStack_2                                     #
#                                                               #
#       Jason Thomas Maynes     - Mark Glover       Jason Lamoureux  #
#       May/June 2002                                           #
#                                                               #
#       This program will find Arg's that H-bond to a guanine in a  #
#       pyrimidine-guanine pair and also are within a cutoff distace  #
#       of the C5 atom of the pyrimidine ring. Will also print out  #
#       statisitics for bond distances that are indicative of a  #
#       loss of pyrimidine-guanine base stacking.
#                                                               #
#################################################################

if ($#ARGV == -1 )
{
        print "\nUsage ./FindArg [file with pdb names - one per line] [H-bond
cutoff distance] [Hydrophobic interaction distance]\n\n";
        exit;
}

print "*****************************************************************\n\n";
print "                        ArgStack_2D                          \n\n";
print "            Jason Thomas Maynes/ Mark Glover / Jason Lamoureux -
May 2002          \n\n";
print "                                                \n\n";
print "*****************************************************************\n\n";

#read in file list
open(PDBFILES,"<$ARGV[0]");
@files=<PDBFILES>;

print "\nYour H-bonding distance: $ARGV[1]\n";
print "Your hydrophobic distance: $ARGV[2]\n\n";

#process each file
for ($currentpdb=0; $currentpdb <= $#files; $currentpdb++)
{
        open(CURRENTPDBFILE,"<$files[$currentpdb]");
        my @currentpdb=<CURRENTPDBFILE>;
        close(CURRENTPDBFILE);
```

```
#extract residues
my @residues;
foreach (@currentpdb)
{
	my @array=split(" ",$_);

	if ($array[0] eq "ATOM" || $array[0] eq "HETATM")
	{
		push(@residues,$_);
	}
	@array=();
}

#look for PyG steps
my $foundPy=0;	#boolean for if currently found pyrimidine
my @foundPyG; #array for coord values for current pyrimidine
my %GHBcoord; #hash for coord values of guanines
my %GSTcoord; #hash for coord values of guanines
my %ARGcoord; #hash for coord values of NH1, NH2, NE of Arg's
my %ARGCzcoord; #hash for coord values of CZ of Arg's
my %residuesdone; #hash to count  no. hb's
my $storePynum; #scalar for current pyrimidine values residue num
my $storePychain; #scalar for current pyrimidine values residue chain
my $foundG=0; #boolean for if found a guanine after a pyrimidine

#elementary FSM for finding Py-G sets
foreach (@residues)
{
	my @array=split(" ",$_);

	my $residue = $array[3];
	my $atom = $array[2];

	my $chain;
	my $resnum;
	my $coordx;
	my $coordy;
	my $coordz;
	#test for residue #'s over 1000
	{
		my $resnumid1 = $array[5];
		my $resnumid2 = sprintf "%.0f",$resnumid1;

		if ($resnumid1 == $resnumid2)
		{
```

```
                        $chain=$array[4];
                        $resnum=$array[5];
                        $coordx=$array[6];
                        $coordy=$array[7];
                        $coordz=$array[8];
                }
                else
                {
                        $chain=substr($array[4],0,1);
                        $resnum=substr($array[4],1);
                        $coordx=$array[5];
                        $coordy=$array[6];
                        $coordz=$array[7];
                }
        }

        #if prev found Py but no G reset
        if (!($residue eq 'G') && !($residue eq "GUA") && $foundPy ==
1)
        {
                $foundPy=0;
        }
        #found a Py
        if ($residue eq 'T' || $residue eq 'C' || $residue eq "CYT" ||
$residue eq "THY")
        {
                $foundPy=1;
                $storePynum=$resnum;
                $storePychain=$chain;
                $foundG=0;
        }
        #if prev found a Py and now a G, store the chain id and residue
# of Py and G
        if (($residue eq 'G' || $residue eq "GUA") && $foundPy == 1 &&
$foundG == 0)
        {
                #ensure that on same chain, also store the coord of
centroid of Py ring
                if ($storePynum < $resnum)
                {
                        push(@foundPyG,"$chain $storePynum
$resnum");
                }
                $foundG=$resnum;
        }
```

```
                #get coord of O6 and N7 for guanine and C2/C5 coord for
Pyrimidine
                if ($foundG)
                {
                        #check for atom, that same chain and that same residue
(continuous G's)
                        if (($atom eq "N7" || $atom eq "O6") && ($storePynum <
$resnum) && $resnum == $foundG)
                        {
                                $GHBcoord{"$chain $resnum $atom"}="$coordx
$coordy $coordz";
                        }
                }
                #get coord of C4 and C5 for guanine
                if ($foundG)
                {
                        #check for atom, that same chain and that same residue
(continuous G's)
                        if (($atom eq "C4" || $atom eq "C5") && ($storePynum <
$resnum) && $resnum == $foundG)
                        {
                                $GSTcoord{"$chain $resnum $atom"}="$coordx
$coordy $coordz";
                        }
                }
                #get coord of all arginine NH1, NH2 or NZ, initialize
residuesdone
                if ($residue eq "ARG")
                {
                        if ($atom eq "NH1" || $atom eq "NH2" || $atom eq "NE")
                        {
                                $ARGcoord{"$chain $resnum $atom"} = "$coordx
$coordy $coordz";
                                $residuesdone{"$chain $resnum"} = 0;
                        }
                        if ($atom eq "CZ")
                        {
                                $ARGCzcoord{"$chain $resnum $atom"} =
"$coordx $coordy $coordz";
                        }
                }

        }

        #calculate the distances from every N* in Args to O6 and N7 in
guanines
```

```
	my @closeArgs; #array to hold chain and id of close guanines and
Arg's
	my %residuesdone; #hash to check if guanine or Arg has been already
included
	foreach $guanine (keys %GHBcoord)
	{
		my @guanarray = split(" ",$GHBcoord{$guanine});
		my $guanx = $guanarray[0];
		my $guany = $guanarray[1];
		my $guanz = $guanarray[2];
		@guanarray = split(" ",$guanine);
		my $guanchain=$guanarray[0];
		my $guannum=$guanarray[1];

		foreach $arg (keys %ARGcoord)
		{
			my @argarray = split(" ",$ARGcoord{$arg});
			my $argx=$argarray[0];
			my $argy=$argarray[1];
			my $argz=$argarray[2];
			@argarray=split(" ",$arg);
			my $argchain=$argarray[0];
			my $argnum=$argarray[1];

			$distance = sqrt(($guanx-$argx)**2+($guany-
$argy)**2+($guanz-$argz)**2);

			if ($distance < $ARGV[1] && ($residuesdone{"$argchain
$argnum"} == 0))
			{
				$residuesdone{"$argchain $argnum"}++;
			}
			if ($distance < $ARGV[1] && ($residuesdone{"$argchain
$argnum"} == 1))
			{
				push(@closeArgs,"$guanchain $guannum
$argchain $argnum");
				$residuesdone{"$argchain $argnum"}++;
			}
			@argarray=();
		}
		@guanarray=();
	}

	#calculate now if successful Arg's are near Py ring, if near, calculate
base stacking distances
```

```
	my @successArgs; #Arg's that have been determined to be close to
guanine, 5'Py, show stacking dist's
	foreach $checkarg (@closeArgs)
	{
		my @argarray = split (" ",$checkarg);
		my $argchain = $argarray[2];
		my $argnum = $argarray[3];
		my $argx;
		my $argy;
		my $argz;
		my $argxNH1;
		my $argyNH1;
		my $argzNH1;


		#get coord of CZ for this Arg
		foreach $argcz (keys %ARGCzcoord)
		{
			my @argczarray = split (" ",$argcz);
			if ($argczarray[0] eq $argchain && $argczarray[1] eq
$argnum)
			{
				my @argczcoordarray = split ("
",$ARGCzcoord{$argcz});
				$argx=$argczcoordarray[0];
				$argy=$argczcoordarray[1];
				$argz=$argczcoordarray[2];

				my $argNH1coord = $ARGcoord{"$argchain
$argnum "."NH1"};
				my @argNH1coordarray=split (" ",$argNH1coord);
				$argxNH1=$argNH1coordarray[0];
				$argyNH1=$argNH1coordarray[1];
				$argzNH1=$argNH1coordarray[2];
			}
		}

		#find C5, C2, O2 of matching pyrimidine, C4/5 of matching G
		my $guanchain = $argarray[0];
		my $guannum = $argarray[1];
		my $pyC5x;
		my $pyC5y;
		my $pyC5z;
		my $pyC2x;
		my $pyC2y;
		my $pyC2z;
```

```
my $pyO2x;
my $pyO2y;
my $pyO2z;

#need to find coord of C5, C2, O2 of Pyrimidine ring
foreach $LINE (@residues)
{
        my @array=split (" ",$LINE);
        my $testchain;
        my $testresnum;
        my $coordx;
        my $coordy;
        my $coordz;
        my $testatom = $array[2];
        #test for residue #'s over 1000
        {
                my $resnumid1 = $array[5];
                my $resnumid2 = sprintf "%.0f",$resnumid1;

                if ($resnumid1 == $resnumid2)
                {
                        $testchain=$array[4];
                        $testresnum=$array[5];
                        $coordx=$array[6];
                        $coordy=$array[7];
                        $coordz=$array[8];
                }
                else
                {
                        $testchain=substr($array[4],0,1);
                        $testresnum=substr($array[4],1);
                        $coordx=$array[5];
                        $coordy=$array[6];
                        $coordz=$array[7];
                }
        }
        if ($testchain eq $guanchain && $testresnum ==
($guannum-1))
        {
                if ($testatom eq "C5")
                {
                        $pyC5x=$coordx;
                        $pyC5y=$coordy;
                        $pyC5z=$coordz;
                }
                if ($testatom eq "C2")
```

```
                    {
                        $pyC2x=$coordx;
                        $pyC2y=$coordy;
                        $pyC2z=$coordz;
                    }
                    if ($testatom eq "O2")
                    {
                        $pyO2x=$coordx;
                        $pyO2y=$coordy;
                        $pyO2z=$coordz;
                    }
                }
            }

            #calculate distance between CZ and C5
            my $distance = sqrt(($pyC5x-$argx)**2+($pyC5y-
$argy)**2+($pyC5z-$argz)**2);

            #calculate vector b/w CZ and NH1
            my $argvecx = $argxNH1-$argx;
            my $argvecy = $argyNH1-$argy;
            my $argvecz = $argzNH1-$argz;

            #calculate vector b/w CZ and centroid of ring
            my $pyvecx = $pyC5x-$argx;
            my $pyvecy = $pyC5y-$argy;
            my $pyvecz = $pyC5z-$argz;

            #calculate distance between CZ and NH1
            my $NH1distance = sqrt(($argxNH1-$argx)**2+($argyNH1-
$argy)**2+($argzNH1-$argz)**2);

            #calculate the angle between the two vectors
            my $angle =
(($argvecx*$pyvecx)+($argvecy*$pyvecy)+($argvecz*$pyvecz))/($NH1distan
ce*$distance);
            $angle = acos($angle);
            $angle = (360*$angle)/(2*3.1415);
            $angle = 90 - $angle;

            #check for cutoff distance and calc stacking distances and save
            if ($distance < $ARGV[2])
            {
               foreach $guanine (keys %GSTcoord)
               {
                   my @guanarray = split(" ",$GSTcoord{$guanine});
```

```
                        my $guanx = $guanarray[0];
                        my $guany = $guanarray[1];
                        my $guanz = $guanarray[2];
                        @guanarray = split(" ",$guanine);
                        my $testchain=$guanarray[0];
                        my $testnum=$guanarray[1];
                        my $testatom=$guanarray[2];

                # calc. c2 - c5 distance, o2 - c5 distance
                        if ($testchain eq $guanchain && $testnum ==
($guannum) && $testatom eq "C5")
                        {
                        $distc2c5 = sqrt(($pyC2x-$guanx)**2+($pyC2y-
$guany)**2+($pyC2z-$guanz)**2);
                        $disto2c5 = sqrt(($pyO2x-$guanx)**2+($pyO2y-
$guany)**2+($pyO2z-$guanz)**2);
                        }
                # calc. o2 - c4 distance
                        if ($testchain eq $guanchain && $testnum ==
($guannum) && $testatom eq "C4")
                        {
                        $disto2c4 = sqrt(($pyO2x-$guanx)**2+($pyO2y-
$guany)**2+($pyO2z-$guanz)**2);
                        }
                }
                        push(@successArgs,"$argchain $argnum $distance
$angle $distc2c5 $disto2c5 $disto2c4");
                }
        }

        $file = $files[$currentpdb];
        chomp $file;
        print "For File $file: \n";

        print "\tAll Args within H-bonding of guanine in Py-G step:\n";
        foreach (@closeArgs)
        {
                my @array=split(" ",$_);
                print "\t\tArg chain: $array[2] residue: $array[3]\n"
        }
        print "\tAll Args within H-bonding and stacking:\n";
        foreach (@successArgs)
        {
                my @array=split(" ",$_);
                my $distance = sprintf "%.2f", $array[2];
                my $angle = sprintf "%0.2f", $array[3];
```

```
                my $d2 = sprintf "%.2f", $array[4];
                my $d3 = sprintf "%.2f", $array[5];
                my $d4 = sprintf "%.2f", $array[6];

                print "\t\tArg chain: $array[0] residue: $array[1] distance CZ to
Py C5: $distance angle: $angle\n";
                print "\t\tStacking: C2-C5: $d2 O2-C4: $d4 O2-C5: $d3\n";
        }
        print "\n";
        #clear all arrays/hashes for next pdb file
        @currentpdb=();
        @residues=();
        @foundPyG=();
        %GHBcoord=();
        %GSTcoord=();
        %ARGcoord=();
        %ARGCzcoord=();
        @closeArgs=();
        @successArgs=();
        @stack=();
        %residuesdone=();
}
```